

E-optimism on a tide of red ink

Telecommunications and semiconductor companies are suffering from a precipitous downturn. But researchers can be sure that a healthy, handy electronic future is on the way.

The author J. P. Donleavy once wrote that electronic books are a bad thing because they cannot be accumulated on shelves: to remind you of your past, to impress your neighbours and colleagues, and to help prevent divorces thanks to the sheer bother of arguing over who owns what. But despite these and other reasons for paper manufacturers' optimism, electronic media have too much going for them not to displace conventional ink on paper for some purposes, at least. The key questions are: how and when?

How, for example, can the portable and visual convenience of printed *Nature* be emulated electronically? Candidate display technologies — electronic ink and paper, organic light-emitting diodes — are already under development, although combining high resolution, high contrast and stability remains a major challenge. Moore's law, still consistent with foreseeable semiconductor technologies, is set to deliver ever-increasing memory and processing power in handy and affordable forms. Hand-held computers and the Wireless Application Protocol are harbingers of the third-generation (3G) technologies that will allow megabits-per-second data transmission and digital media convergence over the air. Thus, a very powerful combination of readability, functionality and portability is coming into sight.

The Internet's infrastructure is also developing apace. Its brute carrying capacity, with fibre optics, optical interconnections and amplifiers and multiple-wavelength transmission, is huge compared with foreseeable needs. Bottlenecks arise, but engineers are developing switches and routers that are faster and more responsive. Beware of those who advocate an intelligent network, however. Far better to have a dumb but flexible network that allows increasingly sophisticated machines, not to mention people, to interact with each other, rather than rely on network operators to (fail to) anticipate our needs.

Value of richness

But even in the absence of friendlier hardware, people are being driven to read clunky screens because of the increasing richness of the medium. The web pages and links that were the very point of Tim Berners-Lee's original vision of hypertext are now easy enough to generate, and fast enough to link to, to be a compelling addition to everyday communication. Streaming technologies allow users to listen to radio stations as they work or, more purposefully, communicate dynamic images and data. For the complex, high-throughput data of some sciences — particle physics, gene-expression arrays, neural imaging and suchlike — electronic media are increasingly the only practical means of communication.

Web applications and services will soon become much easier to access. Several software companies are developing the ability to provide Internet users with unique identities or electronic passports. Online suppliers can agree to recognize those identities just as they can a credit card — and an online purchase will be correspondingly easier as well. Gone, then, will be the days of multiple usernames and passwords. For researchers, this freedom, combined with increasing collaboration between publishers in developing links and specialized searching, will make it ever easier to pursue research strands across the literature.

Web applications are also set to become more computationally powerful. In an initiative known as the Semantic Web, the Internet itself is hoped, within a decade, to become an everyday vehicle for communication between technologies as well as people. Meanwhile, supercomputer centres are turning themselves into nodes of the Grid — the computational equivalent of the Internet, using the Internet itself, in which computer algorithms are automatically redistributed around the network as, from one instant to another, demand and local capacity fluctuate. Companies, such as IBM, national funding agencies and the European Union are all developing Grid initiatives.

Beating the strain

The social aspects should not be forgotten either. Although the global free-for-all spirit is still prevalent among Internet users, a corresponding threat of a tragedy of the commons, in which the resource ultimately collapses under the strain, is likely to be circumvented by regulations and privatization. Less contentiously, Internet investment is rapid in many developing countries, which are generally free of the necessity to pay off previous investments in slower infrastructure, and the Internet's availability is beginning to stimulate a significantly more generous distribution of information.

All very exciting, except that much of this might seem like an increasingly distant and naïve pipe-dream, given the news in recent months. Embarrassing for enthusiasts, geeks and salespeople is the fact that the telecommunications, IT and semiconductor industrial sectors have seen a swathe of cancelled orders, inventories that greatly exceed demand, redundancies and closures. Sales of mobile telephone handsets are in freefall in the West. Telecommunications companies and their suppliers are suffering from the impact of huge debt burdens from 3G licence auctions. This has limited their ability to raise investment and repay loans; as share values have tumbled, dividends have been cancelled and credit ratings reduced to junk.

Nevertheless, and despite the further collapse of the dotcom e-commerce bubble, the underlying picture is positive. Infrastructure continues to develop as governments, businesses and consumers continue to pay for improved network and online services. The convenience, security and value of the more resilient sectors of online commerce — both in retail and business-to-business transactions — continue to grow. Many troubled companies believe it is important to maintain R&D programmes to secure eventual recovery from market downturns. History also suggests that unforeseeable convergence of technologies will lead to new sources of growth and market opportunities. At most, therefore, and barring catastrophes, the arrival of the technologies heralded above should be delayed by the few years of an economic cycle.

For its part, *Nature* will remain committed to the quiescent joys of paper for years to come. But most researchers are literature scavengers as well as grazers, and therefore demand a highly linked and functional electronic literature. *Nature* is committed to their needs too. We, like them, look forward to navigating the waves of hardware and software to come, and damn the stock market's torpedoes. ■





Seeds of doubt
Study on organic
food called into
question
p666



Standard bearer
Wellcome Trust
unveils plans to beat
misconduct
p667



Strong support
Nigeria assumes
leading role in
African science
p668



Caught in a trap
Animal dysfunction
raises doubts over
lab data
p669

Bush compromise raises doubts over stem-cell resilience

Jonathan Knight, San Francisco

At last, US biomedical researchers can exhale. After months of delay, President George W. Bush announced on 9 August that he would allow the government to pay for human embryonic stem-cell research — provided the cells used had been made before that date.

But the Bush compromise, which had been anticipated for more than a month (see *Nature* **412**, 107; 2001), leaves US stem-cell researchers with a steep technical challenge: to prove the promise of stem cells for treating incurable diseases before the permitted cell lines run out.

Although Bush stated that some 60 cell lines would qualify for grants from the National Institutes of Health (NIH), fewer than 12 are in the published literature. Stem-cell researchers were unaware of the existence of so many others until the announcement. "I am mystified myself as to this number of 60 lines," says Roger Pedersen at the University of California, San Francisco.

The NIH, which produced the figure for

Bush, insists that the cells exist. "They are out there and we are confident that we will be able to make them accessible to our investigators," says Lana Skirboll, associate director of science policy at the agency.

Nevertheless, many researchers have expressed doubts as to whether all 60 will be obtainable, or useful. "It's pretty obvious that the two issues are quality and availability," says Douglas Melton, chair of molecular and cellular biology at Harvard University.

Currently, only one company in the United States, WiCell in Wisconsin, distributes human embryonic stem cells. WiCell has five lines, all developed by James Thomson at the University of Wisconsin, which it hopes to make available to university researchers for a flat fee of \$5,000.

Two other companies, one in Australia and the other in Singapore, say they have 10 lines between them to distribute. Details of the remaining lines are not yet available.

But access to the materials may be the least of researchers' worries. More than a decade of experience with mouse stem cells



Government line: George Bush takes advice from top aide Karen Hughes before his speech.

suggests that the lines eventually expire.

Although stem cells divide indefinitely, mutations in their DNA gradually accumulate until the cells no longer function properly, explains Thomas Doetschman, a pioneer in mouse stem cells at the University of Cincinnati. For example, mouse stem cells injected into a mouse embryo differentiate to become part of the adult, but lose this ability after extended growth in culture.

WiCell says that after two years of culturing, Thomson's original lines are unchanged and completely normal. But this may not be true for all of the 60 lines the NIH claims exist.

Some stem-cell lines last longer than others. Today, most mouse stem-cell work is done with five fairly hearty lines, but to end up with these, researchers had to create and sift through hundreds, says Andras Nagy, a stem-cell researcher at Mount Sinai Hospital in Toronto. "If someone had limited the mouse stem-cell research to 60 lines back in 1985, the knowledge and promises in this field would not have been close to what we have now," he says.

But some biologists saw the Bush announcement as a useful step forward. Brigid Hogan, a mouse embryonic stem-cell researcher at Vanderbilt University, Nashville, Tennessee, says: "You can get a lot of work done, and if the data start looking good and there are a lot of exciting results, there will be pressure to derive more cell lines."

Biologists officially welcome plan

Paul Smaglik, Washington

President Bush's announcement permitting limited research on human embryonic stem cells (see above) was greeted, at least officially, with cautious optimism by American biologists.

The American Society for Cell Biology, the Federation of American Societies for Experimental Biology (FASEB) and the Association of American Medical Colleges (AAMC) each issued statements supporting the policy. However, they tempered that support with concerns that the scientific viability and commercial availability of the available lines will slow the pace of discovery.

But they did not officially voice the community's latent fear that the policy will severely constrain US stem-cell research.

Before the policy was announced, Tony Mazzaschi, associate vice-president for

research at the AAMC, stated that such a compromise approach "doesn't make any sense, hold any water, or gain the administration anything". Yet after Bush's speech, the AAMC released a statement saying it was encouraged that funding could go forward. FASEB also said it was pleased the research could go forward, "if only in a limited way".

Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, says he is surprised that the scientific community has not been more direct in its response. Because so many limits have been placed on what can be funded, many universities will be reluctant to allow their scientists to pursue the research, he says. "The president effectively sold a ban. The scientific community has decided to accept what can only be called a crumb."

Nutritionists question study of organic food

David Adam, London

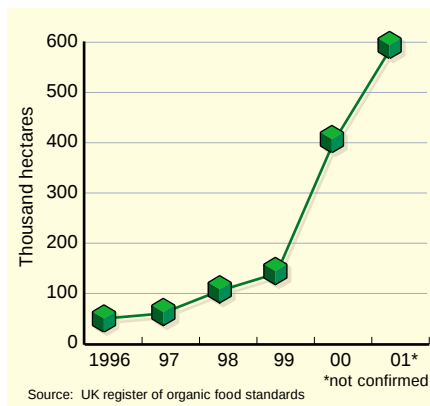
Some food researchers may find the concept hard to swallow, but the British government has pledged to investigate more closely the supposed nutritional benefits of eating organically grown food.

The Food Standards Agency (FSA) says it will host a seminar later this year to consider "how consumer choice between organic and conventional food could be further informed by research". Political pressure has been growing in Britain for some definitive information on differences between the two.

But some food experts have already questioned the value of the exercise, fearing that it will only serve to highlight 'differences' of no real significance to human health. "It seems to me to be quite unlikely that one could do anything reasonably sensible in that area without spending an enormous amount of money," says Eileen Rubery, a public-health specialist formerly with the Department of Health and now at the University of Cambridge.

Sceptics say that direct, comparative studies of organic and non-organic produce are difficult to construct because of extraneous variables such as climate and soil conditions. They also point out that epidemiological studies of people who consume organic and non-organic produce would be expensive and prone to influence by factors such as other lifestyle differences between the two population groups.

The FSA's decision to consider research into the question followed the publication on 6 August of a report by the Bristol-based Soil Association, which was billed by its authors as showing significant differences in the mineral and vitamin contents of organic and non-organic food. The Soil Association is a non-profit organization that promotes organically grown food, which is produced without



Fresh growth: the amount of British farmland under organic cultivation has risen dramatically.

recourse to synthetic fertilizers or pesticides.

The FSA says the new report does not make a convincing case that organic food is safer or more nutritious, but adds that there is enough public interest in the matter to merit further research. Demand for organically grown fruit, vegetables and meat has soared in Britain (and elsewhere in Europe) in the past few years: the UK organic food market is now worth some £600 million (US\$854 million) per year, and the amount of British farmland dedicated to organic farming has risen tenfold since 1996 (see graph).

The image of organic food as a healthy alternative has been boosted in Britain by public scepticism about genetically modified crops, and by the mad cow disease epidemic, which has undermined faith in modern, intensive farming techniques.

The FSA will not elaborate on the kind of research it has in mind, saying only that "a very wide range of options" was being discussed. "It's not possible at this stage to say exactly what that research should be, but it

will have to be something that can establish whether there is or isn't a difference," says a spokesperson for the agency. Some supporters of organic agriculture would like to see long-term human trials, perhaps involving prison populations or people in care homes for the elderly to minimize the effects of environmental factors.

The Soil Association commissioned its report in response to comments made last August by John Krebs, a zoologist and chairman of the FSA. Krebs said of consumers who buy organic produce: "They're not getting value for money, in my opinion and in the opinion of the Food Standards Agency, if they think they're buying food with extra nutritional quality or extra safety."

Soil Association director Patrick Holden claims that the report proves Krebs wrong. "It asserts that there is indicative evidence suggesting nutritional differences between organic and non-organic food," he says.

Holden says the report, prepared by nutritionist Shane Heaton, an independent consultant, is the clearest picture yet of the research carried out so far. Heaton reviewed some 400 previous papers that considered or compared organic and conventional foods, in areas such as mineral and vitamin content, food safety and observed health effects in animals or people. Such reviews have been carried out before, but Heaton says that his study breaks new ground because it excludes many trials that failed to compare the two properly.

For example, of 99 studies that compared the nutrient contents of organic and non-organic foods, Heaton says that only 29 were valid. The rest looked at food that had not been grown according to the strict guidelines laid down by organic groups. Of the legitimate studies, a handful showed significantly higher amounts of minerals, vitamins and dry matter in organic food.

But this is not enough evidence to satisfy other nutritionists. "It's very silly to make claims when so many different factors like the weather, soil and season could result in these changes," says retired food scientist and consultant Ralph Blanchfield. "The old saying that two peas in a pod are alike just doesn't apply; two peas in a pod can still analyse differently."

Some supporters of organic farming also played down the significance of the study, saying that typical consumers of organic food were more likely to benefit from their general eating habits than from extra nutrition in particular foods. "As much as anything it's a question of diet, rather than the individual food items," says Martin Wolfe, an organic farming researcher at Wakelyns Agroforestry, a research farm in Suffolk, UK.

♦ www.soilassociation.org
♦ www.foodstandards.gov.uk



Value for money? Organic food is gaining popularity, but researchers question its touted benefits.

Wellcome Trust sets out fresh misconduct standards

WELLCOME TRUST



Standard bearer: the Wellcome Trust has released draft guidelines for handling scientific misconduct.

Erica Klarreich, London

Research institutes seeking grants from the Wellcome Trust will soon have to adhere to a new set of standards for scientific conduct. And in the absence of government action on research misconduct, observers say, the charitable trust's guidelines are likely to set a *de facto* standard for British universities.

A draft of the guidelines, issued for comment on 26 July, defines good science practice and sets out procedures for handling cases of misconduct, such as plagiarism and falsification of data. But in their current form, critics point out, they do little to protect people who blow the whistle on misconduct.

The trust, which funds nearly 20% of Britain's biomedical research, plans to insist that all institutions receiving its grants adhere to the standards from October 2002.

Robert Terry, a senior policy adviser at the trust, says the peer-review process cannot prevent scientific misconduct. "This issue has been discussed for a while in Britain, but hasn't moved much forward," he says. "We felt we had to make our own position clear."

The Medical Research Council, Britain's main publicly funded biomedical granting agency, expects institutions to investigate allegations concerning their own researchers. The Wellcome Trust says it will carry out its own investigations in "exceptional cases", such as when it feels its reputation is at stake.

The United Kingdom has no national body to investigate misconduct cases, and each of the six research councils, which distribute government grant money to scientists, has its own guidelines for universities.

"This is the biggest thing to happen in the field of scientific misconduct in the United Kingdom for years," says Drummond Rennie, deputy editor of the *Journal of the American Medical Association* and an expert on scientific misconduct in Europe. "I'm impressed by this document's decisiveness, thoughtfulness and clarity."

The extent of the Wellcome Trust's support for biomedical research is likely to turn its guidelines into *de facto* national guidelines, Rennie says. Terry agrees: "I doubt that universities will institute a system that only applies to Trust researchers," he says.

In December 2000, the European Science Foundation encouraged scientific societies to press for the creation of national bodies to investigate misconduct. Denmark is the only European nation that has such a body.

But Marcus De Ville, a spokesman for the UK Office of Science and Technology, says there are no plans to create one in Britain. "It is for the scientific body as a whole to regulate itself," he says. "The peer-review process is the system which verifies the credibility of scientific research, and that process works."

While broadly welcoming the guidelines, Herbert Arst, a professor of microbial genetics at Imperial College in London, says they offer inadequate protection to people who make allegations. Whistle-blowers are often stigmatized in the scientific community and may face expensive libel suits, he says, causing misconduct to go unreported.

Terry responds that the final Wellcome Trust guidelines may include stronger protection for whistle-blowers.

Academies called to task over human cloning débâcle

Laura Bonetta, Washington

The US National Academies has come under fire for inviting researchers who plan to clone human beings to a forum held in Washington on 7 August.

Severino Antinori, director of the International Associated Research Institute for Human Reproduction in Rome, and Panos Zavos, director of the Andrology Institute of America in Kentucky, used the meeting to announce plans to start cloning experiments in November. Television broadcasts subsequently led with the news in both the United States and Europe.

Also present was Brigitte Boisselier, scientific director of Clonaid, a company set up by the Raelian sect, which intends to clone humans.

"The academies should only have the best, most credible, and authoritative scientists to present data," argues David Mangus, a bioethicist at the University of Pennsylvania in Philadelphia.

Alan Trounson, deputy director of the Monash Institute of Reproduction and Development in Clayton, Australia, adds that media attention may help to "attract money and thereafter the skills to conduct this unfortunate experiment".

Debbie Stine, an academies official, defended the invitation, saying that the speakers had scientific credentials and both sides should be represented. But other participants say the would-be cloners gave no details of what they are doing, and no real dialogue took place.

There is also concern that the tidal wave of media coverage blurred the distinction between therapeutic cloning — which many scientists support — and reproductive cloning.

The forum was part of a fact-finding process for an academy panel that will report to Congress next month. The House of Representatives has passed a bill that bans human cloning, and the Senate will soon consider similar legislation. ■



Three of a kind: Severino Antinori (left), Brigitte Boisselier and Panos Zavos.

SHAWN THEW/GETTY IMAGES

Nigeria takes the initiative in African science

Alison Abbott

Nigeria has set up a national scientific advisory council and is planning to help fund the African Academy of Sciences (AAS) — raising hopes that the country will provide badly needed leadership for science in Africa.

The developments follow the 1999 election of President Olusegun Obasanjo, which ended 16 years of military rule in Nigeria.

After the announcement last month of a US\$94-million satellite programme (see *Nature* **412**, 110; 2001), Obasanjo said that Nigeria would donate US\$5 million to a newly created endowment fund to support the AAS, a pan-African organization based in Nairobi, Kenya.

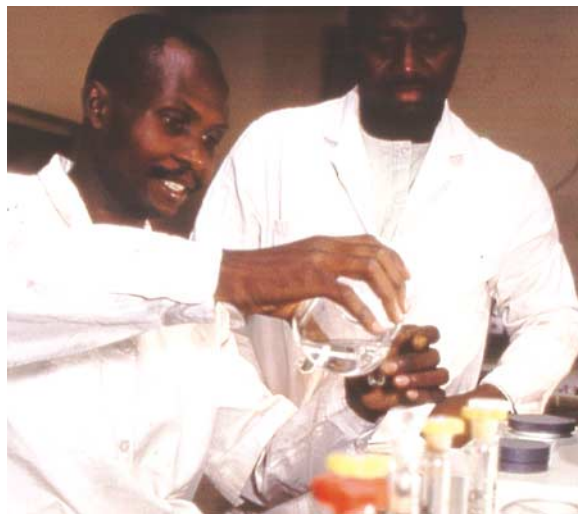
At the same time, his government has inaugurated a new Nigerian science advisory council, comprising seven scientists, including some from outside Nigeria. It is chaired by Mohamed Hassan, a Sudanese mathematics professor, who is also president of the AAS.

Nigerian science and technology minister Turner Isoun says the decision to fund the AAS is “the most strategically enlightened decision the government has made” — mainly on account of the academy’s role in increasing scientific capacity throughout Africa.

But Isoun says that the decision will ultimately be in Nigeria’s own interest. Nigeria has one of the strongest scientific communities in Africa, and is a prime beneficiary of the AAS’s fellowships and programmes. Of the 104 fellowships that the academy has so far distributed to young African scientists, 28



Leading light: Mohamed Hassan, chair of Nigeria's science council, aims to boost African research.



M. ROSE/PANOS PICTURES

were awarded to Nigerians.

Isoun also defends the satellite programme, which had been criticized as over-ambitious. “It is not right that Africa should be ‘sentenced’ to appropriate technology,” he says, arguing that Nigeria should be able to solve its own problems rather than “buying information from other countries”.

At the moment Nigeria spends little on research — although no reliable figure has been published, the total is probably less than 0.5% of gross national product. But the prevailing mood is such that a large spending increase can be expected, Isoun says.

Nigeria is the first country to commit money to the endowment fund for the AAS, which was set up in 1985 to promote science

in Africa. As well as electing members, of which it now has 150, the academy also publishes a journal, *Discovery and Innovation*, and runs projects intended to build scientific capacity in Africa.

Initially, the AAS was supported by donations from large US charities, including the Rockefeller, MacArthur and Carnegie foundations. Now most of its support — around US\$1.5 million per year — comes from the Swedish development agency Sida. “We set up the endowment fund because we need a stable, independent source of financing,” says Hassan. The AAS now plans to lobby South Africa and other African countries for permanent support.

♦ <http://www.oneworld.org/aas>

Feminized fish encourage Japan to test pollution links

David Cyranoski, Tokyo

A comprehensive search for a link between pollutants and the appearance of ‘feminized’ male fish is being considered by Japan’s National Institute for Environmental Sciences (NIES). A decision is expected later this month.

Recent research has shown that certain male fish in the seas around Japan have physiological properties usually found in females. For example, some males produce high levels of vitellogenin, an egg-yolk protein, and others even make oocytes — precursor egg cells — in their testes.

Some researchers suspect that these problems are due to chemical ‘endocrine disrupters’, which at low doses can interfere with hormones. But most of the evidence for this comes from apparent correlations between the presence of feminized fish and heavily polluted areas. Little progress has been made in identifying the molecular



Female empowerment: egg cells developing within the testis of a ‘feminized’ male fish.

mechanisms and chemicals that are to blame.

Some scientists — as well as chemical manufacturers — dispute the existence of the endocrine-disrupter effect, regarding it as an excuse to regulate hundreds of chemicals without evidence that they cause harm.

The fact that some fish naturally switch sex further confounds the problem. “We need caution before jumping to a conclusion

that something is abnormal,” says John Sumpter of Brunel University, UK, who pioneered a study of intersexed roach. “The evidence linking cause and effect is slim.”

But over the past couple of years, Japanese studies have indicated that the effect is widespread. “We are seeing an effect in a variety of fish,” says Shinya Hashimoto, an environmental chemist at Shizuoka University.

Researchers say that a comprehensive study could also help to establish a database for pollution monitoring. “It could be a valuable measure of relative pollution over the areas,” says Takahiro Matsubara of the Fisheries Research Agency in Hokkaido.

The NIES is scheduled to complete a feasibility study this month to determine how best to carry out the proposed investigation, but budget constraints may prevent a more extensive study, officials say.

S. HASHIMOTO

Animal data jeopardized by life behind bars

Jonathan Knight, San Francisco

Research animals raised in standard laboratory cages appear to develop a brain defect that could affect the outcome of experiments, according to behavioural scientists.

The findings, presented last week at the 35th Congress of the International Society for Applied Ethology at the University of California, Davis, are likely to spark debate among neuroscientists about the wisdom of keeping laboratory mice in austere, standard cages.

Cages that provide more stimulating environments for animals could help solve the problem, researchers say. But space constraints and concerns about standardization of experiments have so far kept such innovations out of most animal research facilities.

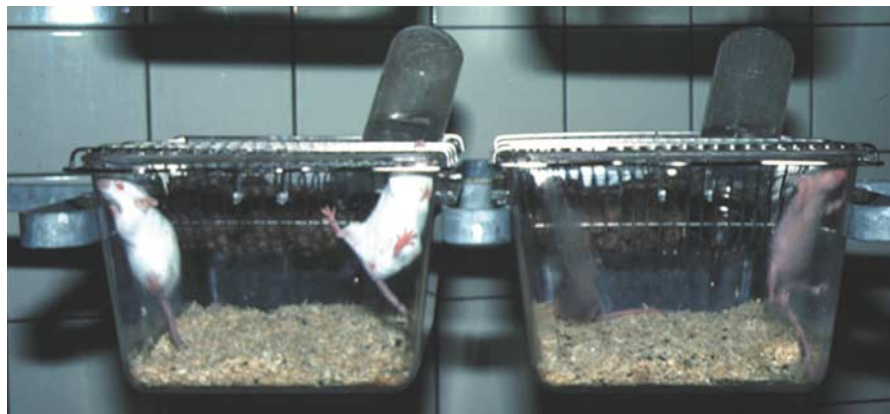
Previous studies established that adding nesting material or littermates to cages increases the number and density of neurons in animals' brains. Rats and mice in enriched cages also perform better in memory tests (see *Nature Rev. Neurosci.* **1**, 191–198; 2000).

Even so, their less astute cohorts reared alone in standard shoebox-sized cages have still been regarded by most researchers as normal. But the new evidence — presented by Joseph Garner, a behavioural scientist at the University of California, Davis — challenges that assumption. Garner's data suggest that the brains of animals housed in barren environments are severely abnormal.

Caged animals often develop repetitive behaviour patterns called stereotypies, Garner said. They pace, groom themselves incessantly, or gnaw the bars of the cage for hours.

In humans, stereotypies are thought to indicate brain damage in a region of the brain known as the basal ganglia, which regulates the initiation of movements. But cage-induced stereotypies have been regarded as merely excessive indulgence in normal behaviour, caused by an abnormal environment.

Garner and his colleagues said that they have now shown that the stereotypies reflect permanent brain dysfunction in parrots. Garner monitored parrot behaviour with a



Stir crazy: mice housed in standard lab cages exhibit a range of abnormal repetitive behaviours.

psychological test normally used to diagnose damage to the basal ganglia in humans. The birds that showed little or no stereotypic behaviour did best on the test, whereas those that spent a lot of time in repeated feather-plucking, for example, failed the test.

Garner believes that the result links the stereotypies to brain defects, and probably applies across species. He has already completed similar studies in voles and is now beginning to work with mice.

Cage-induced stereotypies were not known in rats and mice until 1996, when Hanno Würbel, an animal behaviourist at the Swiss Federal Institute of Technology in Zurich, used an infrared camera to record the behaviour of these nocturnal creatures at night. In the dark, the vast majority of mice engaged in behaviours such as repetitive bar biting and cage scratching. But most people

who work with mice are unaware of the problem, because they only handle the animals in the light.

Even if brain defects turn out to be widespread in cage-reared animals, it is not clear how this will affect research. For many types of experiment, it may matter little. "If the process being studied is fundamental, it should reveal itself over and above these differences," says Fred Gage, a neurobiologist at the Salk Institute in La Jolla, California.

"The stereotypy effects might not matter if they were the same across all labs," says Georgia Mason, a zoologist at the University of Oxford and Garner's graduate adviser. "But what we find is that stereotypy varies quite a bit."

"The irony is that all this barren environment that has been put upon animals for standardization may be the source of the variability," Garner says. ■

Chemistry journal reacts to dispute

David Adam

A chemistry journal plans to deal with an acrimonious dispute over a submitted paper by taking the unusual step of publishing it along with an addendum that disputes some of its conclusions.

Langmuir, which is published by the American Chemical Society, is planning to publish the nanotechnology paper on its website this week. An addendum states that the scientist who supervised the work takes issue with some of its content, its ownership and aspects of how the work was carried out.

The paper, written by Peter Schwartz, a physicist now at California Polytechnic State University, describes a way of patterning very thin lines of DNA onto a gold surface. Schwartz says he developed the technique while working in the laboratory of chemistry professor Chad Mirkin at Northwestern University, Illinois. Schwartz

left Northwestern in August 2000 following several disagreements with Mirkin.

Langmuir accepted the peer-reviewed paper in March this year, but Mirkin effectively blocked its publication by writing to the society asserting that the data were incomplete and that Schwartz was trying to pass off collaborative research as his own. The journal agreed to publish the work only when Schwartz accepted the inclusion of the addendum. "The addendum solution is a compromise, but one that is not great for anyone involved," Mirkin says.

Schwartz — who believes that his results cast doubt on the performance of a different DNA micropatterning technique called 'dip pen lithography', developed by Mirkin's group — says he is "very relieved" that other researchers can now attempt to reproduce his work.

► <http://pubs.acs.org/journals/langd5/index.html>



Russia and United States open airlock to more astrotourists

Washington NASA and Rosaviakosmos, the Russian space agency, have agreed a set of guidelines that will allow more wealthy tourists to visit the International Space Station.

The two agencies quarrelled earlier this year over the visit of American businessman Dennis Tito, who was eventually allowed to spend eight days on the orbiting facility. The new guidelines detail training and other requirements for paying guests, who will travel to the station in the Russian Soyuz capsule.

The Russian space agency is currently training South African Internet entrepreneur Mark Shuttleworth, who, like Tito, will pay upwards of \$20 million for his trip. The French and Italian space agencies also have bought seats on Soyuz for their professional astronauts. Claudie Haigneré of France is scheduled to visit the station in October.

New Zealand servicemen on course for court case

Wellington New Zealand veterans of British nuclear tests in the 1950s have been advised



Explosive issue: veterans of the Pacific nuclear tests are hoping to sue the British government.

by their lawyers that they have legal grounds for suing the British government.

Over 500 New Zealand servicemen were ordered to stand on the decks of two frigates during nine hydrogen-bomb tests in the Pacific (see *Nature* 412, 5; 2001). Veterans claim that their life expectancy has been reduced, and that there is an unusually high rate of genetic disorders among them and their children.

The veterans now intend to raise funds to make a claim in the British High Court. The

group will seek compensation for veterans, widows, and possibly their children.

Japanese institute rebuts spying charge

Tokyo The Institute of Physical and Chemical Research (RIKEN) in Saitama, Japan, was not involved in the industrial espionage alleged by the US Department of Justice in May, according to a team of lawyers commissioned by the institute.

The allegations arose after the United States charged two molecular biologists, Takashi Okamoto and Hiroaki Serizawa, with conspiring to steal DNA samples from the Cleveland Clinic Foundation (CCF), where Okamoto headed a lab until July 1999 (see *Nature* 411, 225; 2001). US officials allege that Okamoto removed the samples shortly before he left the CCF to take up a position at RIKEN.

A report by the lawyers says that there is no evidence that the samples were used at the Japanese institute or that RIKEN intended to acquire the CCF materials by employing Okamoto. Serizawa, who was at the University of Kansas Medical Center, is currently awaiting trial in the United States. Okamoto resigned from RIKEN late last month, and the United States may yet seek his extradition.

Red rain leaves Indian scientists battling demons

New Delhi An unidentified fungus and a meteor shower are the two possible solutions being put forward to explain mysterious red rain experienced by the southern Indian state of Kerala late last month.

Researchers from the Centre for Earth Science Studies in the state capital Thiruvananthapuram say that spores of a fungus found in the rains may be responsible for the red colour. But an earlier theory — that the colouring was due to a meteor storm — has been revived by reports from the local population of a sonic boom.

The red rains, as well as other events, have fuelled rumours that the state is under the spell of a demon. Despite the current monsoon, almost 150 wells there have dried up over the past two months. Researchers from the National Geophysical Research Institute in Hyderabad say that shifts in underground channels after earthquakes earlier this year are the most likely explanation.

Inquiry to learn from blighted British farms

London The role of science in tackling Britain's recent foot-and-mouth epidemic is

set to come under scrutiny as the Royal Society takes up a government request to investigate infectious diseases in livestock.

The inquiry, which is scheduled to last for a year, will bring together veterinary scientists, virologists and epidemiologists, as well as representatives of farming and consumer groups. It will be headed by Brian Follett, former vice-chancellor of the University of Warwick. The panel will examine what can be learned from the recent outbreaks of foot-and-mouth disease and swine fever.

The study brings the total number of foot-and-mouth inquiries under way in Britain to ten. They range from small studies by committees of back-bench members of parliament to a full-scale investigation of the future of British farming.

Museum seeks scientists to experiment on the public

London Scientists who are running short of subjects for their experiments have been invited to use visitors to London's Science Museum in their research.

Project officer Sabiha Foster says the museum is looking for new researchers to take up residence in the purpose-built research area of its Live Science project. All research must be non-invasive and



Leisure lab: museum visitors will provide the raw data for the Live Science area's experiments.

painless and will require approval by a museum ethics panel.

Live Science has already hosted geneticists from University College London, who scanned visitors' faces, and used the data to help identify the genes involved in facial development.

♦ <http://www.sciencemuseum.org.uk>



To boldly go

The scientists on board *HMS Challenger* had little idea what they would find when they sailed from Portsmouth in December 1872. The voyage, the first thorough survey of the world's oceans, is now recognized as the start of modern oceanography. But because so little was known about the oceans at the time, the journey was one of exploration. The scientists simply set out to see what they could find.

In the decades that followed, oceanography left some of this sense of adventure behind. Like a lot of science, it is now built on the careful testing of preformed ideas. But for some, this switch has been premature. A fresh exploratory drive is needed, say researchers, because vast stretches of ocean remain unstudied. "Without exploration, we would continue on a familiar path, with familiar subjects, enjoying an occasional surprise. But with exploration, our purpose is to discover these surprises," says Craig McLean, director of the Office of Ocean Exploration (OOE) at the US National Oceanic and Atmospheric Administration (NOAA) in Silver Spring, Maryland.

Luckily for oceanographers, plans for exploratory missions have caught the imagination of US politicians. Funding has been made available, and the first of a new series of missions is scheduled for September. Add in recent advances in the technology needed to support these missions, and the future looks bright for ocean science. "We are at a turning point," says Sylvia Earle, former chief scientist at NOAA. "Look at what was done for aviation and aerospace during the twentieth century. We're poised right now to parallel that in the present century for the oceans."

Political interest took shape last October, when a panel set up by former President Bill Clinton delivered its plans for ocean exploration. Its report called for US funding agencies, including NOAA and the National

The nineteenth century scientists who studied the oceans were explorers, not hypothesis testers. Mark Schrope finds that modern-day oceanographers want to revive this pioneering spirit.

Science Foundation (NSF), to support new exploration projects. These calls were answered last year when Congress awarded NOAA \$4 million for 2001 to set up the OOE. Congress is currently considering a funding increase to \$14 million for the programme in 2002. Ultimately, researchers hope that support will reach the annual level of \$75 million recommended in the panel's report.

All at sea

Researchers have traditionally found it difficult to persuade funding bodies to support exploratory missions. "How can you write a proposal to the NSF to do something when you don't know what it is you're going to find?" says Robert Ballard, leader of the team that discovered the wreck of the *Titanic*, and founder of the Institute for Exploration, a deep-water archaeology research centre in Mystic, Connecticut. The NSF does have a small budget for exploratory research, but grants cannot exceed \$100,000 and are typically much smaller.



Into the blue: the submersible Alvin (above) will be used on new missions aimed at reviving the exploratory spirit of *Challenger* (top).

The OOE aims to support explorers — defined in the panel's report as people who have "not narrowly designed the observing strategy to test a specific hypothesis". Researchers on OOE missions will naturally have ideas about what they may find and hypotheses they hope to test, but these will not be the driving forces behind the expeditions. Instead, the voyages will be open-ended. The OOE will also fund the development of technologies to advance exploration, such as underwater-imaging technology and crewed or robotic submersibles.

By using interdisciplinary teams of scientists, the new missions should be prepared for surprises. The hope is to make discoveries as exciting as the first glimpses, in 1977, of hydrothermal vents and the astonishing array of life that surrounds them, but to avoid the problems that arose from that mission's nar-

row focus. The expedition consisted solely of geologists who were unprepared for biological work and so ended up preserving samples in “vodka and good bourbon”, says Ballard, who was co-chief scientist on the mission.

Marine biotechnology is one of the disciplines that stands to gain from the new programme. Biologists have been mining the ocean floor, much as they do the rainforests, looking for compounds that could be useful — some of which have already been used to develop new cancer treatments. But for the field to fulfil its potential, researchers need samples from the huge areas of ocean that have never been surveyed.

Studies of biodiversity should also benefit, says David Guggenheim of the Center for Marine Conservation in Washington. Data on the oceans are patchy, so it is hard to assess the impact of human activities on marine life. Deep-water corals are a good example. Because oceanographers have only recently realized how extensive these corals are, they were not aware of the damage being done by fishing vessels trawling the ocean floor for cod and haddock. “Exploration is absolutely critical,” says Guggenheim, “because unless scientists can get a handle on what is in the oceans, they certainly can’t develop practices to protect them, and potentially significant problems could go unnoticed.”

Making a splash

Several existing projects have already been incorporated into the OOE programme, including the Sustainable Seas Expeditions. This project — funded by NOAA in partnership with the National Geographic Society and the Richard and Rhoda Goldman Fund, a charitable organization based in San Francisco — was planned as a series of exploratory missions to marine sanctuaries. It was expanded this year to examine sites of interest in the Caribbean and the waters off South Carolina.

The OOE has also organized a new mission, dubbed Deep East, in near-record time. Set to begin in September, it will focus on three areas off the northeast US coast. Scientists on each of the legs will be drawn from a variety of disciplines and will use the research submersible Alvin to explore specific areas.

The first leg will focus on a region, about 165–300 kilometres off the coast of New England, known as Georges Bank. The team will study the deep-sea corals that grow there as well as the movement of animals between different areas of the bank. Researchers on the second leg will study the biodiversity of the Hudson Canyon off New Jersey.

The final stage will probe the Blake Ridge, about 165 kilometres east of Charleston, South Carolina. There the main interest will be a large tongue of the seabed in about 2,000 metres of water. In 1995, researchers with the Ocean Drilling Program, an international project to investigate what lies beneath the



Sea view: exploration should give scientists a better idea of deep-sea life such as this giant squid.

sea-floor, discovered substantial quantities of methane hydrates in the ridge.

Methane hydrates are deposits of methane trapped in ice crystals. Researchers estimate that worldwide these deposits contain three times more fossil fuel than the world’s oil resources. But because the hydrates are generally spread out over large areas in deep waters, they have not yet been exploited. Deep East will test ideas about how to estimate the extent of hydrate deposits using acoustic surface measurements. The team will also study the site’s geology and palaeontology, as well as the plant and animal life supported by the methane that seeps out.

Next year, as long as its funding gets congressional approval, the OOE will support an expanded programme that will include expeditions to other areas off the US coasts, as well as either Antarctica or the Arctic — regions that are particularly underexplored.

Back to bed

The OOE will also help to fund an emerging form of exploration, using a series of long-term, unmanned observatories. Last year, a US National Academy of Sciences report outlined a plan for a series of seabed observatories to provide data on areas that are not feasible for ships to visit for long periods (see *Nature* **406**, 449; 2000).

The academy report envisaged a network of stations continuously monitoring the biology and chemistry of the surrounding ocean, but it also called for a dynamic element to the project. A series of rapidly deployable ocean buoys could be used to monitor events such as algal blooms and volcanic eruptions. The OOE plans to support existing projects to develop the necessary technology, says McLean.

If the new exploratory programme does get the full funding recommended by the exploration panel, it could have big consequences for oceanography. Marcia McNutt, president of the Monterey Bay Aquarium Research Institute in California, is confident

that some well-established hypotheses will lose their footing when tested in places that no one has ever been. And Andrew Shepard of the National Undersea Research Center at the University of North Carolina at Wilmington expects a host of questions to arise from the research. “It’s the first step in the scientific process,” he says, “see what’s there.”

Despite this confidence, the nature of exploration dictates that the benefits of any one mission are unpredictable. There is always a chance that a trip will fail to turn up anything especially interesting.

But those involved believe the missions will quickly show their worth. For these researchers, the voyages are an attempt to contrive those moments of luck that lead to leaps in understanding. “Anytime you go to a new place with new technology you are bound to find something,” says McNutt, “It doesn’t matter what you find, ultimately it will be put to very good use. History has shown that.” ■

Mark Schroppe is a science writer in Melbourne, Florida.

► <http://oceanexplorer.noaa.gov>



Expect the unexpected: Robert Ballard is searching for surprises in the deep.

The best supporting actors

Glial cells were long thought to play a peripheral role in the theatre of the brain. But some neuroscientists now believe that they are intimately involved in the way the brain processes information. Bas Kast charts the cells' move into the limelight.

Until ten years ago, glial cells seemed destined for a life of anonymity. Despite being the most numerous type of cell in the human brain — outnumbering neurons by ten to one — glia were seen as little more than packing material. For most researchers, it was the chemical and electrical signals used by neurons to communicate that were the key to understanding the brain.

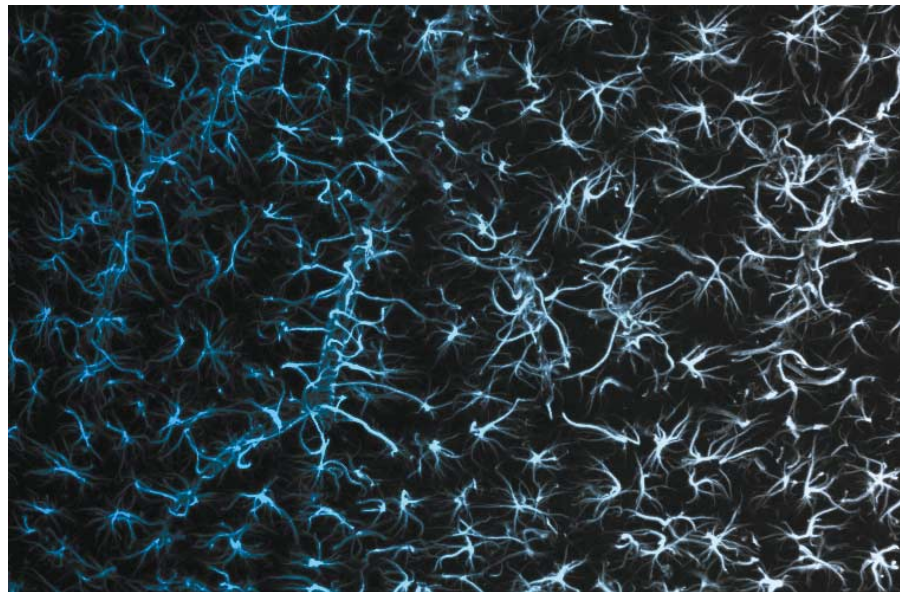
But glial cells are now high on the neuroscience agenda. A series of studies in the 1990s showed that these cells do communicate with neurons, leading some researchers to suggest that our picture of how the brain processes information needs to be revised. For the small band of scientists who have devoted their careers to studying glia, the surge in interest is long overdue. “Finally,” says Maiken Nedergaard, a neurophysiologist at New York Medical College in Valhalla, “glia are starting to receive the attention they deserve.”

Glial cells come in different types, but all were originally thought to play menial roles. One such function, holding brain tissue together, earned the cells their name — glia derives from the Greek word for glue. Some glia wrap themselves around neurons, providing electrical insulation. Others, such as the star-shaped astrocytes that are the focus of much of the new work, have been implicated in a range of jobs. These include supplying neurons with nutrients¹ and maintaining the levels of ions used by neurons in signalling².

Unsung heroes

One of the reasons glial cells have been neglected is that they do not generate electrical pulses. Such voltage pulses are used by neurons as part of their communication system. They are sent along a fibre, called an axon, that projects from the neuron's body. Experimental techniques for studying these pulses revealed little of interest when applied to glia, and the lack of pulses led most researchers to assume that the glia were not involved in information processing.

But electrical pulses are only part of the brain's communication system. Once a pulse reaches the end of an axon, it triggers



Stars of the show: glia such as these astrocytes are being linked to information processing in the brain.

the release of chemicals called neurotransmitters. These diffuse across the synapse — the gap between the axon and the ‘dendrite’, the receiving fibre of the target neuron. Once across, they dock with special proteins, known as receptors, on the dendrite's membrane. Some receptors can react to the arrival of neurotransmitters by opening channels in the dendrite's wall, allowing ions to flow into and out of the cell.

Astrocytes wrap themselves around some types of synapse, and so come into contact with the neurotransmitters that flow across the gap. Evidence dating back to the 1970s shows that glia have receptors for neurotransmitters similar to those on neurons, but neuroscientists had never seen glia use them.

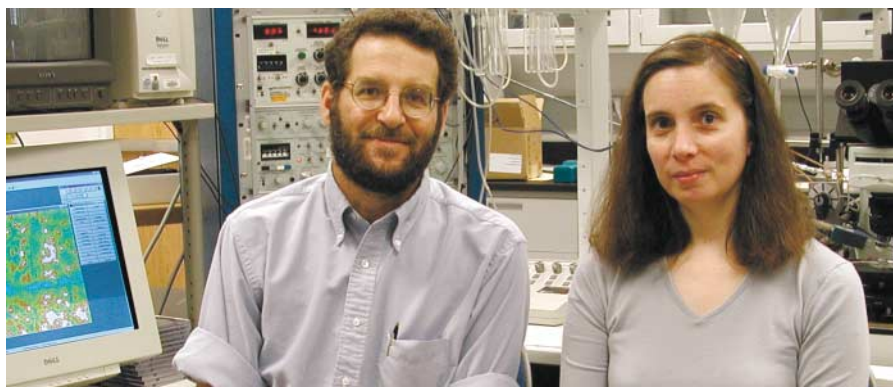
In the late 1980s, research began to reveal that receptors on glial cells could react to glutamate, a common type of neurotransmitter, by opening channels in the glial cell's membrane^{3,4}. In 1990, Ann Cornell-Bell and her colleagues, then at Yale University in New Haven, Connecticut, showed that glial cells in laboratory cultures use their receptors to receive signals from neurons⁵. Cornell-Bell found that the channels opened by glutamate on the glial-cell membrane allow external calcium ions to flow into the cell. At the same time, the cells release calcium ions from internal stores. The combined effect is a rapid increase in the level of calcium ions inside the glial cell.

The Yale team also showed that a rise in calcium levels in one glial cell triggers similar increases in neighbouring cells. These cells, in turn, pass the signal on, causing a wave of rising calcium levels to pass through the local network of glia. Not only were neurons communicating with glial cells, but the glia seemed to be sending signals among themselves.

A series of studies has now shown that the calcium wave is mediated by two different signals⁶. One of these uses gap junctions, areas where two neighbouring glia come into direct contact. The release of calcium ions from the glia's internal stores is controlled by a series of different molecules. These can



Brain power: Maiken Nedergaard speculates that glial cells could be involved in cognition.



Eyes down: Eric Newman and Kathleen Zahs have found evidence that in retinas (right), the calcium waves passing through astrocytes (shown in blue) can influence the way neurons react to light.

flow over gap junctions and trigger the release of calcium ions in adjoining cells. In addition, the rise in calcium levels causes the glia to release adenosine triphosphate (ATP), the molecule used by cells to carry energy. ATP molecules can diffuse across extracellular space and, by opening receptor channels, trigger increases in calcium levels in the glia they come across. This provides a means for the calcium wave to jump between cells that do not make direct contact.

By the early 1990s, many neuroscientists were convinced that neurons could talk to glia, and that glia could talk among themselves. Later that decade, evidence emerged for the final link in the network — signal transmission from glia to neurons.

In 1994, Nedergaard⁷ showed that rising calcium levels in astrocytes were followed by a similar increase in calcium levels in nearby neurons. The signal appears to be transferred across gap junctions between the two cells. The same year, Philip Haydon, then at Iowa State University in Ames, described how glutamate emitted from glia can dock with nearby neurons and raise calcium levels in those cells⁸.

More recently, researchers have started to suspect that the way in which glia wrap themselves around synapses might affect the flow of neurotransmitters. Knut Petter Lehre of the University of Oslo and Dmitri Rusakov of

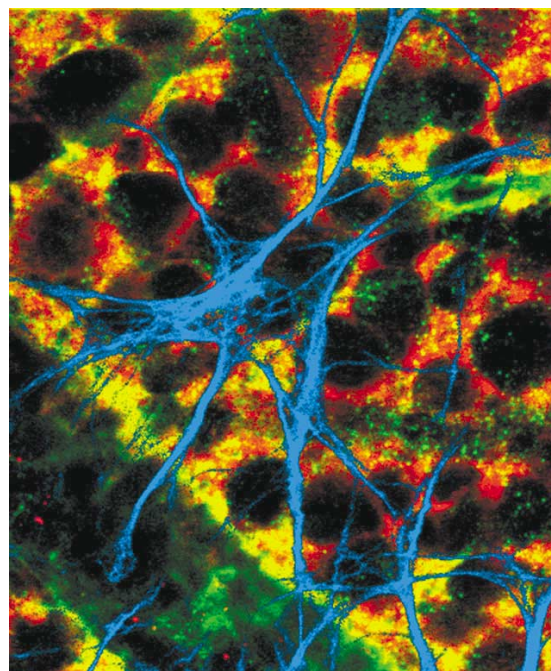
the Institute of Neurology in London are working on a detailed study of the distribution of glia around synapses. Initial results suggest that the glia wrap themselves tightly to the receiving dendrite, but leave more gaps around the transmitting axon.

Dimitri Kullmann, one of Rusakov's colleagues at the Institute of Neurology, suggests that the close association with the dendrite helps to restrict neurotransmitters to their main target. But by leaving gaps around the axon, glia allow the signal to diffuse back around its sides. Here the neurotransmitters can dock with receptors, perhaps providing a feedback loop by which the signalling cell can regulate its own firing.

Three's a crowd

The transformation of neuroscientists' perception of glia has led Haydon, now at the University of Pennsylvania in Philadelphia, to suggest that the classic picture of the synapse needs to be redrawn. He believes that glia surrounding the junction between two neurons should be seen as the third member of a 'tripartite synapse'⁹.

Studies of glial cells in retinas taken from rats support this view. In 1997, Eric Newman and Kathleen Zahs of the University of Minnesota in Minneapolis showed that the calcium wave originally seen in cell culture by Cornell-Bell also occurred in the rat retinas¹⁰.



The following year, the same researchers showed that this calcium wave influences how neurons in retinas respond to light¹¹. The neurons fire when exposed to light, and Newman and Zahs saw that many of the cells fired less frequently as the calcium wave passed. This was just the kind of interaction between glia and neurons that Haydon was proposing.

Later the same year, Richard Robitaille from the University of Montreal in Canada, provided evidence for a three-way synapse in a living animal¹². He studied the neuromuscular junction in frogs. This specialized synapse is the point at which neurons meet muscles, and neurotransmitters released here control muscle movements.

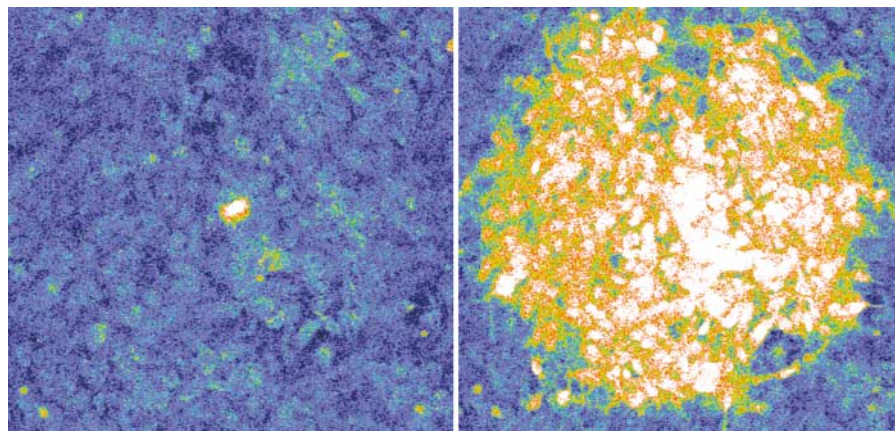
If neurons at the junction are induced to fire repeatedly for long enough, the amount of neurotransmitter released by each pulse starts to decline. Researchers were not sure why this happens, but most assumed that the process was controlled by chemicals in the neurons or the synapse.

Robitaille injected GDPβS, a substance that inactivates chemical messengers known as G-proteins, into the glial cells that surrounded the synapse. G-proteins play a key role in the signalling pathway that leads to increased calcium levels. With the G-proteins disrupted, the glia could not react to neurotransmitters in the normal way.

Robitaille found that the decline in neurotransmitter release stopped when the G-proteins were inactivated. The process seems to be controlled, at least in part, by the glial cells surrounding the synapse. Robitaille speculates that the glia could be acting as a brake on the emission of neurotransmitters, preventing neurons from using up all of their supplies.

Evidence for a tripartite synapse in the mammalian brain comes from studies of

P. ROFUIJE, NEWMAN/K. ZAHs



Surf's up: a calcium wave spreading out from a single cell (left) to hundreds of neighbouring glia.

M. SIMARD/M. NEDERGAARD



Three-way split: Philip Haydon says glial cells are involved in communication at the synapse.

► astrocytes in slices of brain tissue taken from the hippocampus, an area involved in the formation of memories. When one neuron fires repeatedly, the strength of the effect it has on connecting neurons can increase with time. This effect, known as potentiation, is thought to be important in learning and has been observed at synapses in the hippocampus and elsewhere in the brain.

Potential difference

In 1998, Nedergaard and her colleague Jian Kang described how astrocytes can control potentiation at hippocampal synapses¹³. They identified synapses at which repeated firing caused potentiation. They then showed that they could induce potentiation by stimulating astrocytes around the synapse, and prevent normal potentiation by applying a drug to astrocytes that disrupted their functioning.

Other researchers are trying to redraw our picture of the synapse in different ways. Read Montague, a theoretical neuroscientist from the Baylor College of Medicine in Houston, Texas, suggests that by isolating the area around the synapse, glia may be assisting a hitherto undiscovered form of neurotransmission.

Neuroscientists have recently found that voltage pulses sent down the axon can also travel 'backwards', away from the neuron's

The findings have upset the picture of the synapse as a one-to-one connection between neurons.

body and back into its dendrites¹⁴. When the pulse reaches a synapse, it opens channels in the dendrite's membrane, allowing calcium to flow in. By isolating the synapse, glia ensure that these small changes in the number of calcium ions in the gap are not rapidly diluted by calcium flowing in from other areas.

Montague argues that this allows the change in calcium levels in the synapse to be felt by the axon of the neuron on the other side of the junction. Several processes, including the release of neurotransmitters, are influenced by calcium concentrations. So the glia may be allowing neurons to communicate 'backwards' without releasing neurotransmitters.

Montague is now collaborating with his colleagues Dan Johnston and Rob Gereau at Baylor and Michael Friedlander of the University of Alabama at Birmingham to evaluate his ideas experimentally. "To show information processing without neurotransmitters would be a big result," he says, "but these are very difficult experiments to do."

The work of Haydon and Montague provides exciting new angles on how the brain may process information, but it poses a problem for researchers investigating how groups of neurons work together. Traditional models assume that the synapse is a one-to-one connection between two neurons. The new findings have disrupted this simple picture. Glial cells may communicate more slowly than neurons — voltage pulses can travel at up to 100 metres per second, over a million times faster than a calcium wave — but they are still sending information, and no one in the field is certain what the conse-

quences of such transmissions may be.

Newman's work is a case in point. His studies suggest that glia in retinas influence the signals received by the brain's visual-processing areas, but he is at a loss as to what they could mean in terms of how humans or animals actually see. "At this point, we have no idea what this could mean for vision," he says. Nedergaard and Kang face a similar problem. "It's clear that glia can modulate hippocampal synapses, but we still don't have any direct evidence that they actually participate in memory and learning," says Nedergaard.

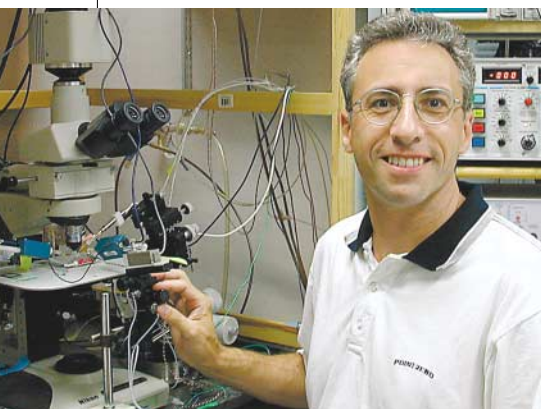
Thought patterns

Comparisons between different species provide a further intriguing hint that glial cells may play a role in cognition. The ten-to-one ratio of glia to neurons in humans drops to one-to-one in rodents¹⁵. And the ratio is reversed in nematodes, with neurons outnumbering glial cells by almost six to one¹⁶. Nedergaard says that the pattern seems to work across all species — the greater the cognitive ability, the higher the ratio of glia to neurons. "It's therefore tempting to think that glial cells play a key role in higher cognitive functions," speculates Nedergaard.

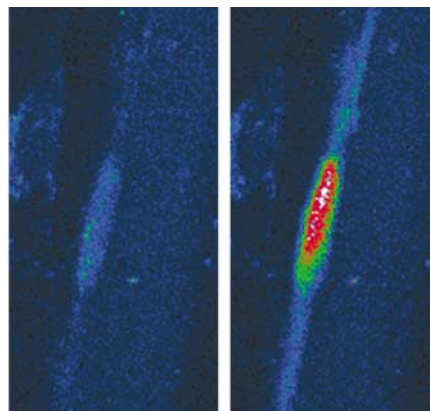
Haydon likens the link between neurons and glia to that between lead actors and their supporting cast and stage hands⁶. The stars could not put on a play alone, and the more complex the performance, the more support is required. "If I think of the spectacular theatre our brains perform every single day, it's probably no coincidence that in our brain the backstage people outnumber the actors by a factor of ten," says Haydon.

Further probing of glial cells' abilities will be needed before neuroscientists can determine the accuracy of Haydon's analogy. But after the past decade's revelations about glial-cell function, the cells are now attracting the attention of a previously uninterested audience of neuroscientists. The backstage crew is finally getting its turn in the limelight. ■

Bas Kast writes for *Der Tagesspiegel* in Berlin.



Message service: by studying the links between neuron activity and the calcium response in glial cells (right), Richard Robitaille has found evidence that glia can exert an influence on neurons.



1. Coles, J. A. & Abbott, N. J. *Trends Neurosci.* **19**, 358–362 (1996).
2. Newman, E. & Reichenbach, A. *Trends Neurosci.* **19**, 307–312 (1996).
3. Usovitz, M. M., Gallo, V. & Cull-Candy, S. G. *Nature* **339**, 380–383 (1989).
4. Marrero, H., Astion, M. L., Coles, J. A. & Orkland, R. K. *Nature* **339**, 378–380 (1989).
5. Cornell-Bell, A. H., Finkbeiner, S. M., Cooper, M. S. & Smith, S. J. *Science* **247**, 470–473 (1990).
6. Haydon, P. G. *Nature Rev. Neurosci.* **2**, 185–193 (2001).
7. Nedergaard, M. *Science* **263**, 1768–1771 (1994).
8. Parpura, V. et al. *Nature* **369**, 744–747 (1994).
9. Araque, A., Parpura, V., Sanzgiri, R. P. & Haydon, P. G. *Trends Neurosci.* **22**, 208–215 (1999).
10. Newman, E. A. & Zahs, K. R. *Science* **275**, 844–847 (1997).
11. Newman, E. A. & Zahs, K. R. *J. Neurosci.* **18**, 4022–4028 (1998).
12. Robitaille, R. *Neuron* **21**, 847–855 (1998).
13. Kang, J., Jiang, L., Goldman, S. A. & Nedergaard, M. *Nature Neurosci.* **1**, 683–692 (1998).
14. Stuart, G., Spruston, N., Sakmann, B. & Häusser, M. *Trends Neurosci.* **20**, 125–131 (1997).
15. Kandel, E. R. in *Principles of Neural Science* (eds Kandel, E. R., Schwartz, J. H. & Jessel, T. M.) 18–32 (Elsevier, New York, 1991).
16. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. *Dev. Biol.* **100**, 64–119 (1983).

Organic movement reveals a shift in the social position of science

Since BSE, the public is less inclined to trust experts.

Sir — Anthony Trewavas will have ruffled a few green feathers in his Commentary “Urban myths of organic farming” (*Nature* **410**, 409–410; 2001), in which he discusses the myths of organic agriculture. Many of the points he raises are valid, and will be accepted by many (though not all) adherents of organic farming.

Trewavas takes a highly positivistic attitude towards science, in contrast to most of the organic movement. In the positivistic world, something is true if established by the falsification of a null hypothesis. By contrast, many supporters of organic agriculture doubt that an objective, proven scientific truth can exist. They believe an agricultural system is more than the sum of its parts, that the reductionistic view widely used in natural science does not see the big picture, and hence that it fails as a political or social tool of analysis. Organic farming, of course, has a strong social and political agenda.

Take, for example, food quality. Trewavas states that “hundreds of rigorous tests have failed to reveal better-tasting properties or improved nutritional value, but have consistently shown that organic produce has lower nitrate and protein content”. The literature he cites, however, raises the point that a lot of food-quality studies concerning the impact of different growing methods lack standards for important parameters such as varieties, growing locations, maturity and storage conditions. These have important effects on compounds influencing the taste and nutritional value of fruit and vegetables.

In addition, only a few studies were performed for certain food groups, leaving the effects of organic or non-organic growing conditions unclear. Furthermore, people are not necessarily buying organic produce because they are unaware of any drawbacks. They may have other reasons which the researchers do not appreciate.

Because the positivistic point of view

doubts what has not been ‘proved’, it will automatically be sceptical towards non-rigorous, anecdotal reported differences between organically and conventionally produced food. On the other hand, many supporters of organic agriculture rely on personal experiences and beliefs that make them more receptive to the idea that there is a difference between organic and conventionally produced food.

In essence, Trewavas is using organic agriculture as a case study of the tendency in modern western society for scepticism about science. Although, as scientists, we may deplore the fact that people are swayed by non-scientific views, the fact is that a lot of them are. Despite the arguments presented by Trewavas, many people believe that organic production systems produce better food, care more for animal welfare and are kinder to the environment. As inquisitive scientists, we should be asking why this is the case.

We should try not to see organic farming in “is it true?” black-and-white terms, but rather look at the social factors leading people to choose organic products. The role of science here is to assist social dialogue rather than simply to deliver the technical ‘truth’. For instance, scientists could identify agricultural and environmental conditions under which organic farming might usefully be practised, and those in which it would not be beneficial.

In the era of BSE and of GM foods, we cannot escape the conclusion that the social position of science is changing. The crucial question is whether dialogue is possible between strict rationalists and those scientists more able to see their subject in a broad social context. That is the wider issue highlighted by Trewavas, and the one that needs to be considered by the whole scientific community.

Annette Mørkeberg, John R. Porter

The Royal Veterinary and Agricultural University, Bülowsvej 17, DK-1870 Frederiksberg, Denmark

Politics defeats science at environment agency

Sir — Legislation intended to “boost science” at the US Environmental Protection Agency (EPA) — which was discussed in your News story “Congress hears plan to boost science at environment agency” (*Nature* **411**, 405; 2001) — is

being aimed at the wrong target.

The agency’s new deputy head would primarily be responsible for “coordinating the EPA’s research portfolio”, but its management of research, while dismal, is less a problem than its regulatory programmes, where policies seem determined more by environmental politics than environmental science.

Adherence to scientific principles in the

formulation of policy has for a long time been alien to the EPA’s corporate culture. An analysis by Resources for the Future, a Washington DC-based environmental think tank (M. R. Powell *Science at EPA*, Resources for the Future, Washington DC, 1999), concluded after an investigation of eight major programmes: “EPA for a variety of reasons is unwilling, unable, and unequipped to address and acknowledge the uncertainties in the underlying science.”

This analysis echoes the conclusions of an expert panel that was commissioned ten years ago by William Reilly, then EPA administrator (*Safeguarding the Future: Credible Science, Credible Decisions*. The Report of the Expert Panel on the Role of Science at EPA. EPA document 600/9-91/050, March 1992)

Fixing the EPA will require far more sweeping and fundamental changes than those currently being proposed. These could range from the creation of an ombudsman panel with the power to impose sanctions on EPA officials who are responsible for unscientific and flawed policies, to dismantling the EPA and redistributing its few essential functions to less scientifically challenged agencies.

Henry I. Miller

Hoover Institution, Stanford University, Stanford, California 94305-6010, USA

Cooperation among labs is appreciated

Sir — I would like to clarify my earlier Correspondence (“Patent dispute hangs over kringle 5”, *Nature* **407**, 128; 2000) about a patent dispute over an angiogenesis inhibitor involving my laboratory and Abbott laboratories.

My laboratory did not contact Miguel Llinás of Carnegie-Mellon University about supplying kringle-containing plasminogen fragments until some months after our angiostatin paper appeared (*Cell* **79**, 315–328; 1994) in October 1994.

Llinás and Johann Schaller of the University of Bern, who have done pioneering work on the biophysical and functional characteristics of the kringle domains of plasminogen, subsequently were very helpful in supplying samples of fragments from human plasminogen to Yihai Cao in my laboratory.

Cao later published two papers on his results, including Llinás and Schaller as authors because of their help in supplying plasminogen fragments and advice.

Judah Folkman

Children’s Hospital, 300 Longwood Avenue, Boston, Massachusetts 0215, USA

A quintessential pluralist

The life of a scientist, wartime strategist and adviser to government.

Solly Zuckerman: A Scientist out of the Ordinary

by John Peyton

John Murray: 2001. 252 pp. £22.50

John Maddox

That Solly Zuckerman, who died in 1993, had more influence on British governments than any science adviser before or since is not disputed. Whether he was a good science adviser is another matter.

A South African Jew who sought fulfilment in Britain in the 1930s, Solly (Lord Zuckerman from 1971) enmeshed himself in a wide range of government business — both important (the choice of bomber targets towards the end of the Second World War, for example) and less so (the role of British badgers in the transmission of bovine tuberculosis). He also ran a successful academic department at the University of Birmingham as well as the complicated business of the London Zoo. He was the quintessential pluralist, the kind of man for whom biographers yearn.

John Peyton's book is not that biography. Rather, it is a memoir, and an idolatrous one at that: the first and last chapters are paeans of unrestrained and even sickly praise. Although the main body of the text frequently refers to Solly's "critics" or "detractors", it nowhere specifies what the criticisms or detractions may have been, let alone the identity of those who may have held such irreverent opinions. Sadly for Peyton (formerly a Conservative politician and minister, now also a lord), Solly was more interesting than a paragon of virtue: he deserves better even from his friends and admirers (among whom I was one).

Solly was, above all, quick-witted: he cottoned on in no time to what people were saying, however inarticulately. He was intellectually and physically energetic. He was gregarious and convivial, which often made him fun. He was also a snob (a possibility that Peyton entertains only to dismiss) in the sense that he gave preferential weight to the opinions of the great and the influential, thereby making powerful friends.

He would fulsomely congratulate for their sagacity those with whose opinions he happened to agree, which may account for Peyton's frequent references to his charm. (Solly's put-downs, on the other hand, were routinely biting.) He knew instinctively that a pluralist could not succeed without helpers and seems to have made it a condition of his many employments that there should be many such to manage his affairs. He was capable of labyrinthine deviousness. Like the



Friends in high places: Solly, seen here with Harold Macmillan, had a way of making powerful friends.

politicians he served, he knew instinctively that power brings the gift of patronage, but he was fickle in the distribution of such benefits to his own loyal and often put-upon servants.

Solly was involved with (and on the side of the angels in) two public issues of huge importance. On the basis of objective studies of the damage caused by German bombs dropped on British cities, he concluded that long-distance bombing of German cities would contribute nothing to the Anglo-American project to invade the European mainland in 1944. With the vital backing of Air Marshal Arthur Tedder, he won his point that it would be better to attack French and German railway networks. Second, with the backing of Lord Louis Mountbatten, he made a speech in 1962 to the NATO council in which he argued that the doctrine of mutually assured destruction (a.k.a. 'MAD' — based on the concept that neither the United States nor its enemies would ever start a nuclear war because the other side would retaliate massively and unacceptably) was a folly ("there would be no winners, only losers") and secured the policy of "graduated deterrence" instead.

He was scientific adviser at the Ministry of Defence during the period (1959–66) when it began to dawn on people that Britain was committed to a range of avionics projects so technically ambitious and expensive that they could not all succeed. (One did, in the shape of the Harrier vertical take-off aircraft.)

Peyton tells us little of Solly's advice on these topics, although he concurs with the accounts of Denis Healey (then defence secretary) and of Solly himself (in his autobiog-

raphy) that his separate briefing of the prime minister about the follies of one of the projects led to an angry falling-out between the minister and his chief adviser. Mere mortals would expect to be fired by the one (Healey) who paid their salary. Solly rose to grander things, becoming scientific adviser to the British government as a whole within two years. Peyton's account of Solly's uncomprehending reaction to the dispute is the best part of the book.

For Solly, that is when the guilt disappeared from the gingerbread. Being closer to the centre of government means being detached from real problems. Solly's solution was a Central Advisory Council for Science and Technology (of which he was chairman), intended to coordinate inter-departmental spending. What he might have done, but did not, was to endow the British government with a durable system of scientific advice across departments. That is a task with which his successors are still struggling.

Sadly, by then (1972), Solly had no substantial constituency in the research community, on which the Rothschild recipe was about to burst. The idea here was to transfer a portion of each research council's budget to an appropriate 'customer' department of government and to require that all departments should be equipped with advice enabling them to be efficient customers.

As if to make amends, Solly threw his energy into the development of the University of East Anglia and especially its School of Environmental Studies, in which he took great pride. Not long before his death, he heard of and was hurt by the dismantling of

his many innovations at the anatomy department at Birmingham.

Peyton sympathizes. Otherwise, he tells us little of Solly's private life, although the narrative makes it plain that Solly put his career before his marriage, one of his prenuptial demands of his wife Joan, the granddaughter of the politician Rufus Isaacs, the first Marquess of Reading, and Alfred Mond, the founder of ICI and the first Lord Melchett. (The "no children" clause was fortunately voided.) In a nice touch, Peyton dedicates his book to "Joan, Solly's wife and the heroine of this story". She, sadly, is also now dead.

John Maddox is emeritus editor of *Nature*.

Signing on the genetic line

Your Genetic Destiny: Know Your Genes, Secure Your Health, Save Your Life

by Aubrey Milunsky

Perseus: 2001. 432 pp. \$27.50, £19.99

Future Perfect: Confronting Decisions about Genetics

by Lori B. Andrews

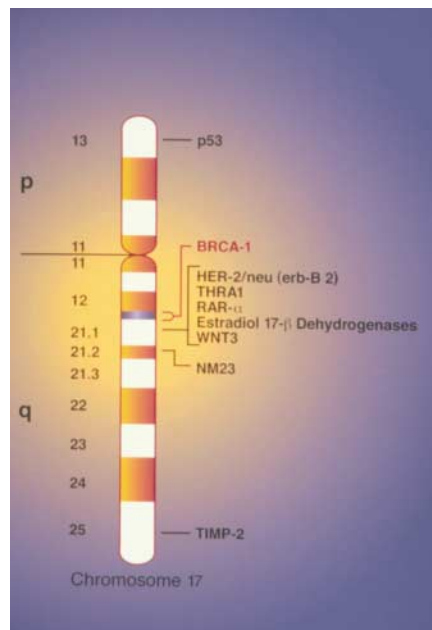
Columbia University Press: 2001. 288 pp. £24.95, £16.95

Ruth Chadwick

Self-help books are nothing new. Self-help genetics, however, may be. *Your Genetic Destiny* is very like many other manuals of health advice. But this one promises that, by knowing your genetic background and predispositions, "you can exercise control over your genetic destiny, secure your health, and ... save your life". Among the details of how to accomplish this is advice on how to compile a family tree. This is said to be one of the most important steps in saving not only one's own life but also those of one's loved ones.

In taking this line, Aubrey Milunsky stands very firmly on one side of the debate about whether it is a good or a bad thing for a person to have information about their genetic make-up and the moral considerations relating to the choice between knowing and not knowing. These moral considerations clearly involve the good of others as well as personal responsibility for one's own health. Milunsky speaks of a "moral imperative" to tell, for example, if you "harbor a certain genetic defect".

To help the reader understand the defects that might exist, Milunsky provides a very useful overview of information about genetics. The coverage includes chromosomes and chromosome aberrations, genes and genetic disorders, X-linked disorders and susceptibility genes, and birth defects and how to



BRCA-1, a gene linked to susceptibility to breast cancer, has been tracked to chromosome 17.

avoid them. He then discusses a number of particular conditions and their genetic basis. These include heart disease, high and low blood pressure, diabetes, cancer, mental illness, mood disorders and Alzheimer's disease. There is also a chapter on the genetic basis of ageing and longevity, reflecting the increasing interest in these topics.

After the encyclopaedic treatment of the genetic basis of different kinds of disorders and conditions, Milunsky offers further practical advice on prevention and treatment, pregnancy, counselling and testing. The penultimate chapter deals with ethical and legal issues. But the premise of the book, as already indicated, is that genetic knowledge is unequivocally desirable and valuable. This is a strong line to take, and one that is widely contested. In *Future Perfect*, Lori Andrews provides an opposing view.

Andrews shows that the 'benefits' of genetic testing raise their own problems, and she points to the danger of hubris in exaggerated claims for what genetics can deliver. She discusses how genetics may have various effects — on the individual (on self-image and reproductive decisions), on groups (particular issues for women, ethnic groups and people with disabilities) and on institutions (concerning health, insurance and employment).

Although all these issues have been discussed elsewhere, *Future Perfect* is most interesting for its explicit focus on the need to find the most appropriate conceptual model for assessing the new genetics. Andrews covers three possibilities: medical, public-health and fundamental-rights models.

One danger she sees in the medical model

— according to which physicians are gatekeepers for genetic services — is that medical personnel may put pressure on people to use genetic services without due regard either for the kinds of impact outlined above or, indeed, for the fact that the information may potentially be inaccurate. The significant issue of the limitations of genetic data as 'information' is one that hardly figures in Milunsky's treatment but which looms large for Andrews.

On the other hand, Andrews does not advocate a public-health model — which would involve attempting to provide improvements in public health through public education about genetic technologies. She provides telling arguments against its appropriateness for genetics. These include the fact that there is a key distinction between a disease that affects many people and a threat to public health, which is typically an imminently dangerous disorder. Andrews also argues that the public-health model provides no standards for what can be considered appropriate testing in genetics. She concludes that the fundamental-rights model is superior; in this model, participation in genetic testing is voluntary and participants retain control over their genetic information.

Milunsky could perhaps be regarded as advocating the medical model that Andrews regards as insufficient. Against this, it could be argued that, because Milunsky's is a self-



DNA dilemma: is detailed knowledge about our genes unequivocally desirable?

help book, there is no place in it for Andrews' concerns about the medical model, such as inappropriate pressure by health-care professionals on patients to undergo genetic testing. Nevertheless, the style in which Milunsky's book is written does raise other concerns raised by Andrews, such as the dangers of overstating the benefits of knowing one's genetic predispositions and susceptibilities. Is it really reasonable to expect that, by following Milunsky's advice, informative though it is, you will save your life?

Both books are very readable and are based on wide knowledge of the literature. Milunsky has user-friendly charts and diagrams; Andrews' book is written in a conversational style. Although Andrews offers a conceptual framework for assessing claims about genetics, her work is set in the same socio-cultural context as Milunsky's. Other conceptual frameworks are therefore possible which might be less individualistic than the fundamental-rights model. By adopting a more socially based position, where individuals have responsibilities as well as rights, it might be possible to find a position between the public-health and individual-rights models. There may be moral reasons, arising out of a sense of social solidarity, for example, why individuals should agree to participate in screening programmes and population-genetic research, provided that adequate protections are in place. ■

Ruth Chadwick is at the Institute for Environment, Philosophy and Public Policy, Lancaster University, Lancaster LA1 4YG, UK.

Physics for non-physicists

Hidden Unity in Nature's Laws

by John C. Taylor

Cambridge University Press: 2001. 490 pp.
£15.95, \$24.95 (pbk)

Peter Landsberg

Are you lost in modern theoretical physics? Do you need help with understanding squarks, selectrons, photinos? Or do you require guidance on the subject of the Higgs particle? Perhaps you need a hint as to the meaning of quintessence in cosmology? If any of these apply, then this is the book for you. But it is rather voluminous, and with good reason. It not only deals with modern ideas, but also with well-established, older theories such as newtonian gravitation, the origin of quantum theory and relativity. It is down-to-earth, however, so there is nothing on the origin of life, tired light or quantum computing.

The distinguished author describes his book as offering "a non-technical tour through the principles of physics". This is a



Wonders of the shallows

The anemone *Sagartia elegans*, above, is found on rocky shores and in rock pools along the Atlantic, North Sea and Channel coasts of northern Europe down to the Mediterranean. It is just one of the diverse species featured in *Photographic Guide to the Sea & Shore Life of*

Britain & North-west Europe by Ray Gibson, Benedict Hextall and Alex Rogers (Oxford University Press, £15.95, \$29.95). The book marks the start of a new series of photographic guides to aid the identification and promote the conservation of Europe's plants and creatures.

good description, but one might add two provisos. First, he delves to some extent into string theory, discusses quantum gravity, penetrates to Grassmann variables, and so on. Although these topics are well explained, it would clearly be misleading to call them 'non-technical'. Second, experimental questions of physics are largely omitted, so a better description of the book would be 'a semi-technical survey of current theoretical physics'.

Regarded in this light, the book is an undoubted success. John Taylor does not try to exaggerate results or make unsupported claims, and he attempts at all times to elucidate complicated matters in simple language. When one deals with superspace, quaternions and the like, this is of course almost

impossible, but Taylor at least has a go.

He also gives some interesting, although not novel, historical insights, including Arthur Eddington's expedition to check on the deflection of light by the Sun and a summary of what various contemporaries thought of Niels Bohr. In addition, Chapter 1 gives a good account of the development of gravitation theory leading to Newton's work.

There are some remarks about the theory of black holes as well as cosmology, and the singularities that occur in these contexts. I expected Taylor to assert at this stage that singularities are normally unphysical, and so cannot exist in nature. However he does not quite take this rather radical step. In the same vein, he observes that "people have often wondered whether the idea of a true point (for particles or charges) was realistic" and so, justifying an extra dimension, he introduces strings to his readers. But a reader may well think that a one-dimensional object is still too idealized. Taylor anticipates this, for he points out that using higher dimensions would lead to rather intractable mathematics.

He also speaks repeatedly of "the universe", when he considers results of calculations. These calculations apply to models and not necessarily to the real Universe, so it would be better to refer to a 'model universe' in such cases. My final quibble refers to the assertion that "entropy should never decrease". But it does, of course, for all the materials in a refrigerator!

I know from personal experience that an author likes to see their book read by all qualified people, and indeed many people will be

For reference

The Penguin Desk Encyclopedia of Science and Mathematics

by Bryan Bunch & Jenny Tesar
Viking Penguin, \$40

The Oxford Companion to the Earth

edited by Paul L. Hancock
Oxford University Press, £39.50

A Dictionary of the History of Science

by Anton Sebastian
Parthenon, £38, \$65

Encyclopedia of Genetics

edited by Eric C. R. Reeve
Fitzroy Dearborn, £105

qualified to read this book. One reason is that the equations used are rather simple. But even simple equations are a barrier for those who have not done much mathematics, and this imposes an unfortunate restriction on the readership. This is a pity, for many worthwhile ideas are expounded here which even a newcomer to physics could understand. I strongly recommend this book. ■

Peter Landsberg is emeritus professor in the Faculty of Mathematical Studies, University of Southampton, Southampton SO17 1BJ, UK. His most recent book is *Seeking Ultimates* (Institute of Physics Publishing, 1999).

The genetic complexity of life

The Misunderstood Gene

by Michel Morange (translated by Matthew Cobb)

Harvard University Press: 2001. 212 pp.
\$24.95, £16.95

Brian Charlesworth

According to the preface of this book, which is a translation of a work in French originally entitled *La Part des Gènes* (Odile Jacob, 1998), Morange's aim is "to put forward a new vision of genes and their function based on recent results from molecular biology". He goes on to assert that "the concept of the gene that is used both by the general public and by many scientists is completely outmoded". By this, he is referring to the role of genes in determining protein structure.

His statement will come as a surprise to most geneticists, as this concept of primary gene function has long since become a truism. Morange seems to feel, however, that the nature and functions of genes are widely misunderstood (he is excluding professional geneticists, presumably), and that this misunderstanding contributes to a widespread hostility to genetic determinism. At one point, he even refers to "the opponents of molecular biology and genetics". These opponents are vaguely defined as people who attribute the complexity of life to components of organisms other than genes, and their views are not discussed in any detail. I felt, therefore, as though I was listening to only one side of an argument; and it is not very clear with whom the argument is being conducted, and why it started.

Most of what Morange has to say is, however, sensible and instructive, and he does an excellent job of presenting molecular, cellular and developmental aspects of genetics to readers who are not geneticists, but who have a fairly good knowledge of biology. Morange's main point is that the product of a single gene is only one part of a complex web of

interacting proteins that results in the structure, functioning and behaviour of an organism. This is, in fact, an old idea, clearly stated in the writings of developmental geneticists such as C. H. Waddington and Sewall Wright in the 1940s and 1950s. The difference now is that molecular genetics has provided detailed evidence about the nature of some of the key players in the game and exactly how they interact with one another. Oddly enough, Morange devotes only a couple of pages to the way in which early embryonic development is controlled by genes, despite the breathtaking advances that have occurred in this field over the past 20 years.

Morange develops various important implications of the broader view of gene action, which he illustrates with concrete examples, mostly from human and mouse genetics. The first is that a change in the structure of a gene, as a result of a mutation or (nowadays) a 'knockout' experiment, may or may not have a drastic effect on the organism, depending on the place of the gene's product in the web of interacting gene products and the ability of the web to compensate for such a change.

Other implications are that multiple, and often surprising, complexes of changes in the organism's characteristics may arise from a single mutation; the nature or even the occurrence of changes may also depend on the nature of the environment, or on the state of other genes. When there is variation in a trait among people, as in their susceptibility to a given disease, a given gene may contribute only a minor part of the variation, and non-genetic factors frequently also con-

tribute to the variability. This applies especially to complex traits, such as behavioural characteristics and lifespan. Again, this all forms part of the stock-in-trade of classical developmental and quantitative genetics.

So genes do not have a one-to-one relation to the characters they affect, and any crude version of genetic determinism fails. This does not undermine the fundamental role of genes in encoding the sequences of amino acids in the polypeptide chains to which they correspond, and which, in turn, determine the three-dimensional structure and biochemical specificities of the folded proteins. Because only the genes are transmissible across generations, only they are subject to the forces of evolution. In a very real sense, therefore, Morange concludes, genes are the ultimate determinants of the development and evolution of organisms.

Little of this is new or controversial, but is probably worth presenting in some detail to show the general reader exactly what is meant by the genetic control of development and behaviour. The book concludes with a discussion of human evolution and eugenics. Morange is rightly cautious about the prospects for widespread genetic manipulation of human genetic material, although positive about limited interventions such as somatic gene therapy and selective abortion of individuals at high risk of genetic disease. Overall, the book is well written and accurate. ■

Brian Charlesworth is at the Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK.

New in paperback

Mystery of Mysteries: Is Evolution a Social Construction?

by Michael Ruse

Harvard University Press, \$19.95, £13.95

Primeval Soup

by Christopher Wills & Jeffrey Bada

Perseus, \$17

"The authors provide an excellent and compelling written overview... this book provides a highly readable survey of the historical prelude to the study of the origins of life." William J. Hagan *Nature* **407**, 451–452 (2000)

The Settlement of the Americas: A New Prehistory

by Thomas Dillehay

Basic Books, \$18

Cradle of Life: The Discovery of Earth's Earliest Fossils

by J. William Schopf

Princeton University Press, \$17.95, £11.95

The Natural Philosophy of James Clerk Maxwell

by Peter M. Harman

Cambridge University Press, £18.95, \$29.95

The Sun in the Church: Cathedrals As Solar Observations

by J. L. Heilbron

Harvard University Press, \$18.95, £12.95

"... John Heilbron's book tells a gripping story with a splendid literary flair... Heilbron tells it in a masterfully human way... with his usual brilliant irony." George V. Coyne *Nature* **402**, 579–580 (1999)

You Can't Eat GNP: Economics as if Ecology Mattered

by Eric A. Davidson

Perseus Publishing, \$16

Color Vision: From Genes to Perception

edited by Karl R. Gegenfurtner &

Lindsay T. Sharpe

Cambridge University Press, £42.95, \$59.95

Stranger than fiction

The tale of the epic voyage made to establish the metric system is an intriguing and exciting one.

Julyan Cartwright

Who thinks about science during a vacation on the Balearic Islands? Majorca, Minorca, Ibiza and Formentera are holiday destinations *par excellence*, not places where famous scientists of the past lived and worked, or where great experiments were carried out. But there was a moment in history when the Balearic Islands were crucial for a major scientific undertaking.

It was 1806, and the French Bureau des Longitudes had the task of determining the Paris meridian — the line of longitude passing through Paris. At the time there was no universally agreed prime meridian; it wasn't until 1884 that an international congress decided it should be the one that passes through Greenwich. The reason for determining the meridian with precision wasn't map-making but the metric system. A few years earlier, French revolutionaries had decided to come up with a new, rational system of measurement. The metre, the new base unit of length, was to be defined as one ten-millionth of the distance from Pole to Equator along the Paris meridian. So it was very important to know exactly how long this meridian was.

A first series of measurements, from Dunkirk to Barcelona, came up with a provisional value for the length and a prototype metre bar was produced. Two young physicists, François Arago and Jean-Baptiste Biot, were given the task of checking and improving the accuracy of this result by extending the observations to the Balearic Islands. Arago and Biot had to light a bonfire on a mountain peak and attempt to view this at night from another peak to obtain their triangulation readings.

South from Paris, the meridian first passes Barcelona on the coast of the Spanish mainland and then runs close to Majorca, more than 200 km away across the Mediterranean Sea. This was too great a distance to make direct observations, so Biot and Arago took triangulations from mountain tops down the Spanish coast to Denia, then across the shorter stretch of sea from there to Ibiza and Formentera — still a difficult task with visual observations — and then finally to Majorca. On Ibiza they used the mountain Camp Vey, and on Formentera, the highest point on the island, La Mola. At the end of 1807, Biot returned to Paris with the observations from Ibiza and Formentera, and left

Arago alone to complete the final readings from Majorca. On Majorca, Arago chose S'Eslop, a peak on the northwest coast, as his viewpoint of Ibiza and Formentera. He had a hut built on the summit and settled in with his instruments for the final series of measurements. But events didn't go according to plan.

War broke out between France and Spain in June 1808, while Arago was on the summit of S'Eslop. Soon Majorcans were commenting that the nightly bonfires were signals and that Arago must be a French spy, and a detachment of soldiers was sent up the mountain to capture him. Arago got wind of this; in his memoirs, he recounts what happened next: "We set off for Palma and we encountered the troops that had come to look

for me. Nobody recognized me because I spoke Mallorquin perfectly. I urged the platoon to continue on the path, and we continued on our route towards the city." (Arago spoke Mallorquin, a version of Catalan, as he was born in the French Pyrenees, a Catalan-speaking region of France.) His escape was only temporary, though — he eventually ended up in Bellver Castle, overlooking Palma de Majorca. Now a tourist attraction, it was then a prison.

In the end, Arago managed to persuade the authorities that he wasn't a spy, and left the island for Algiers. From there he took a ship headed for Marseilles, but his bad luck continued — the ship was intercepted by Spanish pirates and escorted to Catalonia, where he was again imprisoned. Once more, he got himself released, and again set sail for Marseilles. This time it was not pirates but bad weather that intervened — it was now December 1808. The ship was forced by a storm to land at a small port in Algeria and was unable to make the winter crossing to Marseilles, so Arago headed back overland to Algiers. Here he was held captive again, this time as a hostage by the Algerians until the French paid for goods sent to France.

This was resolved in July 1809, and after a year-long odyssey, Arago finally arrived back in France to file his scientific reports, and



Troubled trip: during François Arago's year-long journey home from the Balearic Islands, he was accused of spying, held hostage, and imprisoned on several occasions.

MARY EVANS

received a triumphal reception in Paris. Arago and Biot's labours confirmed the accuracy of the original measurements; in the end, the prototype metre differed from the original meridian definition by just 0.02%.

Arago became an eminent physicist, making many important discoveries in optics (Arago's spot) and electromagnetism (Arago's disc). He later entered politics, becoming Minister of War and the Navy, and he abolished slavery in French territories. Biot too became a physicist of note; his best-known contribution is the Biot-Savart law in electromagnetism.

Today, Arago is honoured in Paris with the Arago monument: a trail of plaques let into the pavements of the city to mark the path of the meridian he plotted. His adventures in the Balearics were fictionalized by his friend Jules Verne, who mentions these episodes in his novels set in the Balearic Islands, *Clovis Dardentor* and *Hector Servadac* (*Off on a Comet*, in its English translation). On Formentera, this fiction is remembered more clearly than the reality, because on La Mola stands a monument — not to Arago and Biot, but to Verne, who probably never visited the islands. ■

Julyan Cartwright is at the Laboratory for Crystallographic Studies, CSIC, E-18071 Granada, Spain.

A purgative mastery

G. J. V. Nossal

The mammalian immune system possesses extraordinary recognition capabilities. On the one hand, it can react to virtually any molecule in the Universe; on the other, it does not normally attack the body itself. The T (thymus-derived) and B (bone-marrow-derived) lymphocyte cells, which have the task of recognition and reaction, somehow seem to know the immunological self. How do they do it? What prompts them, for example, to attack a kidney transplant so ferociously that strong drugs have to be administered to keep them at bay, whereas they leave the individual's own kidney alone?

What is the molecular basis of the immunological self? The issue is rendered more complex by apparent exceptions to self-recognition. Healthy people can have antibodies in their serum that react with bodily components, and some autoreactive T cells are also present. Indeed, some have questioned the centrality of self–nonself discrimination and, as a passionate believer in its importance, I welcome the chance to revisit the subject in the molecular era.

Three key concepts underlie these apparent paradoxes. First, the receptors on the surface of lymphocytes have not been created to react with particular pathogenic microbes or, for that matter, with particular molecules on foreign kidney cells. Rather, the whole universe of receptors together constitutes a repertoire, one or more elements of which are capable of recognizing

any foreign molecule. The immune system works on selectionist, not instructionist, lines. This means that the affinity of binding between antigen and lymphocyte receptor will vary over a wide range. Below a particular threshold value, an encounter between a lymphocyte and an antigen will not engender a signal that is sufficient to trigger that cell into activity.

Second, as lymphocytes mature within the thymus or bone marrow, they go through a developmental stage during which specific receptors first appear on the cell surface. At that particular checkpoint in the lymphocyte's history, contact with self antigens may lead to apoptosis (programmed cell death). I termed this process 'clonal abortion' to distinguish it from 'clonal selection', namely the cascade of activation events in more mature lymphocytes which initiate immune responses. Perhaps fortunately, this rather infelicitous phrase has been replaced by the term 'negative selection'. As a result of such negative selection, lymphocyte populations that leave their birthplace in the thymus or bone marrow to join the circulation have been purged of their most dangerous self-reactive elements.

Third, the lower-affinity anti-self cells will not be triggered into activity simply by meeting self antigens. Lymphocyte activation requires co-stimulatory signals delivered either by a scavenger cell or another lymphocyte. In fact, there are also inhibitory signalling pathways. There are numerous receptors on lymphocytes for ligands that are

Immunological self

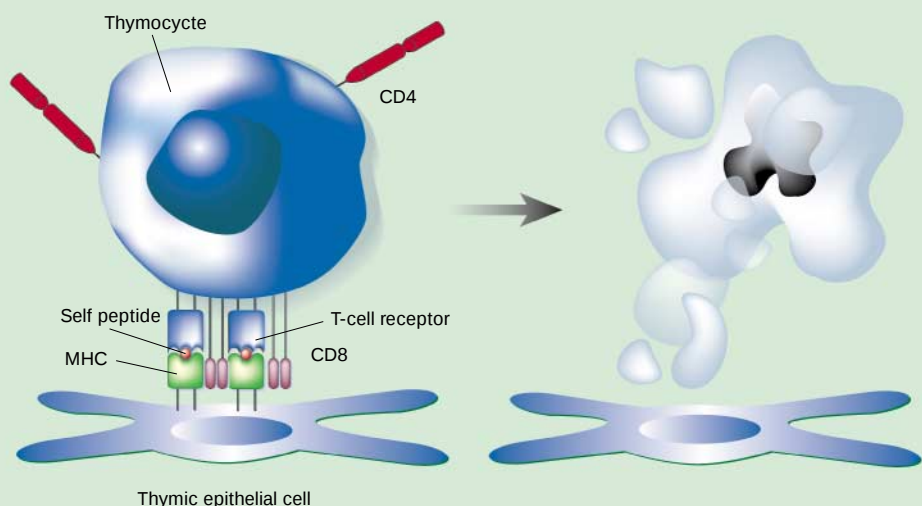
An extraordinarily precise mechanism of self-recognition holds us back from the brink of autoimmunity.

either on other immune cells or work as soluble, secreted molecules that modulate antigen-delivered signals. More of these ligand–receptor pairs and their downstream signal-transducing pathways are being discovered by the month. The lymphocyte constantly integrates this stream of signals, and the mere existence of a self-reactive cell need not imply autoimmunity.

Mix and match

How are lymphocyte receptors generated? The genes that encode them are assembled from tandemly arrayed series of minigenes, which are brought together during the development of each lymphocyte through a unique process of somatic translocation. Each receptor, when finally formed, is the product of five minigenes selected from scores or hundreds. Even further diversity is fostered through the addition or deletion of tiny portions of the minigenes where they join up. Each cell is allowed to have only one receptor specificity, and thus the repertoire of receptors is really a repertoire of unipotent lymphocyte cells.

B cells make antibody molecules, which recognize three-dimensional configurations on the surface (usually) of protein or carbohydrate molecules; for this population, the product of the whole genome constitutes the immunological self. Wherever there are genetic differences — polymorphisms — between individuals, immunological selves will differ. Self antigens in the marrow, such as membrane proteins, are especially important in provoking negative selection by apoptosis. Interestingly, cell death is not instantaneous: the putatively 'bad' anti-self cell is given a brief reprieve, during which it is



Negative selection: the T-cell receptor of a thymocyte binds too strongly to a self peptide and major histocompatibility complex on a thymic epithelial cell, a move that spells death for the thymocyte.

Autoimmune diseases, as a group, constitute the third-greatest burden for clinical treatment.

Active mechanisms ensure that the circulatory system brings many proteins to the thymus for processing and presentation, to encourage the suicide of thymic self-reactive cells.

allowed another attempt at translocation of immunoglobulin genes, in case this results in a non-self-reactive receptor. This process is called 'receptor editing'. It was my good fortune to discover that soluble self antigens in the bone marrow deliver a weaker negative signal, leading not to cell death but rather to a refractory state in which the B cell, on leaving the marrow with a full complement of receptors, is less easily triggered. I called this 'clonal anergy', and on this occasion the term stuck! Surprisingly low concentrations of antigen can induce anergy.

Activated B cells initially make antibodies that are identical in specificity to the assembled immunoglobulin genes, but upon repeated antigenic stimulation the antibodies get better — how? This process of 'affinity maturation' occurs in a specialized structure called the germinal centre, a highly mutagenic environment in which B cells multiply. Here their immunoglobulin genes mutate, allowing the selection of variants that fit better with the antigen in question. Cells that mutate within the germinal centre and happen to acquire anti-self reactivity rapidly undergo apoptosis, another safeguard against autoimmune B cells.

Getting a complex

T lymphocytes cause local inflammation, kill virus-infected cells and help to activate B cells. The T-cell receptor is antibody-like and is also the product of minigene translocations. T cells recognize short peptides tethered to the floor of a groove in a specialized self molecule called the major histocompatibility complex (MHC). If a given T-cell receptor is altogether the wrong shape for the MHC, that T cell is of no use whatsoever. Why design such a complex system? Many pathogens go 'underground', hiding inside cells where antibodies cannot find them. Specialized organelles and enzyme systems exist that fragment intracellular pathogens (such as viruses), associate their peptides with the MHC, and display the complex on the cell surface. In this tricky way, T cells get to see the 'insides' of germs.

Furthermore, peptides from the cell's own molecules are cycled onto the cell surface together with the MHC, possibly so that cells with mutated proteins can be eliminated by the vigilant T cells. So, for the T-cell universe, the immunological self consists of the MHC (which comes in various forms)

plus a myriad of self peptides in the grooves of the MHC molecules.

The thymus therefore faces a double task of selection. It must get rid of T cells that do not recognize the MHC at all while positively selecting those with a receptor that recognizes the MHC. However, it cannot allow the emergence of cells with high affinity for the MHC as these would be autodestructive, and thus it negatively selects cells with strong binding to the MHC plus the myriad self peptides in the groove. To ensure that self-reactivity is rigorously censored, the thymus itself synthesizes many (perhaps most) self proteins, including ones that are not directly related to thymus function.

Active mechanisms ensure that the circulatory system brings many proteins to the thymus for processing and presentation, to encourage the suicide of thymic self-reactive cells. There is now conclusive evidence that the sole discriminator between positive and negative selection is affinity of binding. The fact that nature should choose such a 'soft' discriminator to account for self versus non-self recognition surprised me greatly. As a result, it is clear that plenty of T cells with some, but not much, anti-self potential will emerge from the thymus. Given the importance of co-stimulatory signals mediated by receptors other than the specific, recognizing one, these low-affinity anti-self cells constitute something of a time bomb. For example, if a virus were to infect a tissue in which the self antigen was present, inflammatory cytokines could propel the T cell concerned into action.

MHC genes are highly polymorphic, and the immunological self is particular and idiosyncratic, reflecting an individual's MHC genes and the entire genome, as represented in the peptides situated in MHC grooves. The symmetrical and learned beauty of the T-cell system is to accommodate itself by recognizing this self in every single T lymphocyte, while purging the dangerous cells with high self-affinity. The great MHC polymorphism relates to the survival advantage for a whole species if at least some members of it have MHC grooves that can fit the key antigenic peptides of a particularly nasty or tricky virus.

Smashing the system

Complex mechanisms often break down. Thus, autoimmune diseases together constitute the third-greatest clinical burden,

after cardiovascular diseases and cancer. For example, defects in thymic physiology can leave areas of the organ bereft of the particular cells responsible for negative selection, allowing autoimmune cells to emerge and escape. There are strains of rats in which negative selection is absent at birth but present at 4 weeks of age. Too late! T cells seeded early from the thymus will induce autoimmunity. Sequestered antigens, which do not fully engage in the negative selection process, can be released through infectious damage, adding autoimmune injury to pathology that might otherwise have been resolved. Microbes can possess molecules that mimic certain self antigens, inducing an appropriate, anti-foreign immune response that causes collateral autoimmune damage by accident.

As a matter of fact, examples of proliferation of autoimmune cells are sufficiently commonplace to demand ancillary mechanisms of silencing of autoimmune clones. One such mechanism is antigen-induced T-cell apoptosis, whereby T cells commit suicide if continuously and repeatedly stimulated by an antigen of which they simply cannot get rid. This will be the case not only for self antigens but also for some chronic infections such as tuberculosis and malaria. Some authors argue that T-cell self-recognition is all that matters, and that autoantibodies made by intolerant B cells are purely markers of autoimmunity and have no pathogenetic significance. I strenuously oppose this view, as high-affinity autoantibodies, for example against the surface of red blood cells or against a hormone receptor, can be devastating. Clearly, both sorts of lymphocyte must be kept on the straight and narrow.

So the immunological self exerts its purgative mastery on the lymphocytes only to a degree. This means that with the multiplicity of low-affinity, potentially self-reactive lymphocytes left in our bodies, we are all constantly teetering on the brink of autoimmune disease. Our lymphocytes are not truly resting; rather they are integrating a wealth of positive and negative signals that arise from their antigen receptors and from many other signal-transducing molecules. Most of the time, like good schoolchildren, their rebellious urges are kept in check. But relatively minor defects or insults can perturb the equilibrium, with disastrous results. ■

G. J. V. Nossal is in the Department of Pathology, University of Melbourne, Victoria 3010, Australia.

FURTHER READING

Kamradt, T. & Mitchison, N. A. N. *Engl. J. Med.* **344**, 655–664 (2001).
Miller, J. F. A. P. *Transplantation* (in the press).
Schwartz, R. H. in *Fundamental Immunology* 4th edn (ed. Paul, W. E.) 701–740 (Lippincott-Raven, Philadelphia, 1999).

Quantum ripples in chaos

Andreas Albrecht

The differences between quantum and classical chaos show up on the smallest of scales. Although tiny, these differences have implications for our understanding of quantum mechanics.

Amere 100 years ago — a recent event in the history of human endeavour — mankind discovered that underlying all the familiar laws of physics is the strange world of quantum mechanics. Superficially, at least, the quantum world appears to contradict many things we intuitively ‘know’ to be true about nature. For example, in the quantum world, position and momentum cannot both be precisely specified. Solid particles have a wave-like nature that allows them to produce interference patterns like ripples on a pond. It is even possible for familiar objects to be assigned strange ‘coherent superposition states’ (for example, to be simultaneously dead and alive in the case of Schrödinger’s famous cat).

Such is the contrast between ‘classical’ physics and the quantum world that even Albert Einstein was unwilling to fully embrace quantum mechanics. Yet those who have carefully studied quantum theory have learned many of its tricks. We have discovered how, under the right conditions, both position and momentum can be well specified. And we know that the classical world is rich with natural decoherence processes that rapidly destabilize most Schrödinger cat states (see ref. 1 for a review). Moreover, Schrödinger cat states have been created in the laboratory using real systems, such as the current flowing in a superconductor^{2,3}. Not quite a cat, but not bad. By now, many aspects of quantum mechanics (both mundane and exotic) have been put to the test, and quantum mechanics has passed with flying colours. Still, there are exciting new phenomena to be explored, especially whenever the quantum and classical worlds overlap. On page 712 of this issue⁴ Wojciech Zurek tackles one of these — the difference between classical and quantum chaos — with some surprising results.

Chaotic behaviour is well understood from a classical perspective, and is typically discussed in the context of a mathematical ‘phase space’, in which there are dimensions for both position, x , and momentum, p . A particle at a given instant can be specified as a point in classical phase space, and the time development of the particle describes a curve or trajectory in phase space. In chaotic systems, particles that start out in virtually identical states (that is, at very close points in phase space) rapidly evolve into completely different states (that is, distant parts of phase space).

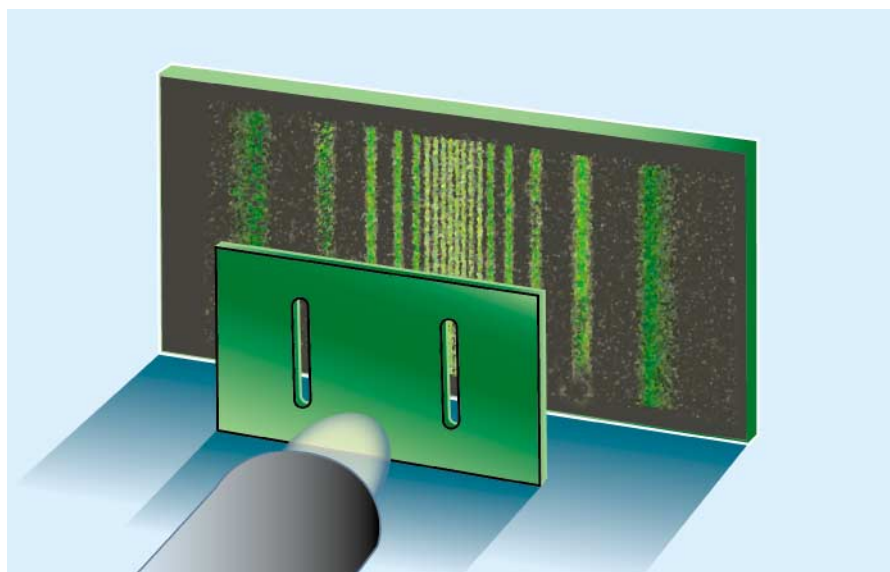


Figure 1 The double-slit experiment. In this, slits are used to create two coherent beams of particles that interfere to produce a pattern on the screen. The coherent beams are a special case of the double-peaked ‘Schrödinger cat states’ discussed by Zurek⁴. The interference pattern has a characteristic size set by a combination of the distance to the screen, the separation of the slits and the momentum of the particles in the beam. The relevant momentum uncertainty is set by the width of the slits. These quantities can be arranged so that the scale of the interference fringes is much smaller than the minimum position uncertainty given by the uncertainty principle. But, as Zurek shows for the general case, there is no contradiction with quantum theory because the uncertainty in position used in the uncertainty principle refers to the width of the entire pattern, not the size of a single fringe. These sub-Planck-scale fringes have physical significance and are related to the sensitivity of quantum states to quantum decoherence.

Because nothing is ever measured with absolute precision, one can never realistically talk about ‘points’ in phase space. Instead, every point (x,p) in phase space is typically assigned a probability, $P(x,p)$. For a well-specified particle this probability peaks sharply at a localized point in phase space. For an ordinary classical object, such as a single billiard ball, a phase-space probability distribution that starts out sharply peaked will remain peaked over time; a small uncertainty in the starting point results in a similarly small degree of ignorance at a later time.

Chaotic systems are dramatically different. A sharply peaked initial distribution gets torn apart by the chaotic evolution, as neighbouring phase-space trajectories rapidly head off in different directions (see Fig. 1d on page 713). A small amount of ignorance at the beginning rapidly translates into huge uncertainties later on, as the distribution becomes highly delocalized.

So how does a chaotic system behave when a full quantum calculation is made? Zurek⁴ gives clear answers to this question, and in the process debunks some widely held beliefs on the subject. He uses the Wigner formulation of quantum mechanics⁵, which gives the state of a particle using the Wigner function, $W(x,p)$. Under suitable conditions, $W(x,p)$ can be interpreted as the classical $P(x,p)$, but W is a more general object that incorporates the full range of quantum behaviour. (For example, W can be negative, whereas P is always positive.)

Heisenberg’s famous ‘uncertainty principle’ of quantum mechanics says that $\hbar/2$ is the minimum value of the product of uncertainties in position and momentum (where $\hbar$ is Planck’s constant). There is a widely held belief that the uncertainty principle requires W to have no features below the ‘Planck scale’ — that is, no features in phase space with an area smaller than $\hbar/2$ ($\approx 10^{-34}$ J s). Zurek shows this belief to be completely false. Along

with a nice general analysis, he gives specific examples of how the Wigner function can have features on scales well below the Planck scale. Zurek shows that there is a smallest scale for structure in W that is set by interference effects, not the uncertainty principle. More importantly, he shows that these features are physically significant: they represent the susceptibility of quantum states to decohering processes.

The idea that quantum systems may have physical features below the Planck scale can be illustrated by the classic example of the double-slit experiment (Fig. 1). In this experiment, a pair of slits creates two coherent particle beams that interfere to produce striking interference fringes. The scale of the interference fringes can be arranged to be much smaller than the minimum uncertainty given by the uncertainty principle. But there is no contradiction with the laws of quantum mechanics because the uncertainty in position used in the uncertainty principle refers to the width of the entire pattern, not the size of a single fringe.

Zurek⁴ shows how a generalized version of these fringes can appear in the Wigner function — typically as ripples on sub-Planck scales — when coherent parts of a quantum state interfere with one another. These fringes are present whenever there is quantum interference. Zurek shows that in the special case of isolated chaotic systems the fringes are essentially always present and emerge very rapidly. Even states that might initially look like localized classical states

are quickly pulled apart by the chaotic evolution into pieces that interfere with one another (see Fig. 1a–c on page 713). Zurek then points out that there are physical effects that can destroy these fringes. This destruction is perpetrated not by the uncertainty principle, but by the coherence-destroying effects of interactions with other physical systems (that is, with the environment). This decoherence is a generalization of another well-known aspect of the double-slit experiment. If the particle beams interact with a measurement apparatus or some other part of the environment in a way that is different for each slit, the coherence is destroyed and the interference fringes disappear.

The struggle to adapt our intuition and insight to the quantum world has been quite an adventure, leading to such creations as transistors, Bose–Einstein condensates and the idea of quantum computation. It is clear that this adventure is far from over, and I expect many more remarkable developments to come. Zurek's article clears away some old misunderstandings and helps us develop a better quantum intuition, which we will need in this exciting future. ■

Andreas Albrecht is in the Department of Physics, University of California, Davis, California 95616, USA.

e-mail: albrecht@physics.ucdavis.edu

1. Zurek, W. H. <http://arxiv.org/abs/quant-ph/0105127>
2. van der Wal *et al.* *Science* **290**, 773–777 (2000).
3. Friedman, J. R. *et al.* *Nature* **406**, 43–46 (2000).
4. Zurek, W. *Nature* **412**, 712–717 (2001).
5. Hillery, M., O'Connell, R. F., Scully, M. O. & Wigner, E. P. *Phys. Rep.* **106**, 121–167 (1984).

wind tunnel^{2,3}. The introduction of smoke or dust streams into the tunnel allows researchers to observe how wing movements deflect oncoming air, and offered a first look at the vortices produced in the insects' wake. These data, combined with detailed analyses of wing kinematics in freely flying insects⁴, provided a basis for evaluating theories about the aerodynamics of insect flight⁵. But it is extremely difficult to obtain repeatable data using live insects, and their small size complicates any effort to quantify airflow.

Against this background, five years ago Ellington *et al.*⁶ published an influential paper showing that the insect wing supports a particular type of vortex, the leading-edge vortex. This is a region of rapidly circulating air, found near the front (leading) edge of the wing, with a low-pressure core. This vortex is stable during the wing's downstroke and might enhance lift, perhaps in part explaining how insects can generate surprisingly large lift forces. The authors were able to describe this phenomenon in detail because they used a mechanical model of a hawk-moth (the 'flapper') with a wingspan of over a metre, which allowed repeatable observations of airflow at a large scale. By injecting smoke directly along the wing's leading edge, the authors revealed that the leading-edge vortex had a helical structure.

Birch and Dickinson¹ have taken this approach considerably further. First, their dynamically scaled model fruitfly (robofly; Fig. 1) has two 19-centimetre-long clear plastic wings whose motion can be precisely controlled. The model is immersed in a large vat of mineral oil, making it much easier to quantify fluid flow over the wing using the technique of digital particle image velocimetry (DPIV) — an increasingly popular tool for studying the mechanics of animal locomotion in fluids^{7–9}. By seeding the mineral oil with small air bubbles and illuminating a two-dimensional slice with a pulsed sheet of laser light, the movement of fluid above and below the wing and in its wake can be quantified with precision. DPIV obviates the need for creative interpretation of smoke trails. Furthermore, the light sheet can be repositioned along the length of the wing to construct a complete three-dimensional picture of flow.

Second, small force sensors at the base of one of the wings (where it joins the fly's body) make it possible to measure the forces perpendicular and parallel to the wing as it flaps, at the same time that DPIV data are acquired. Third, the wing can be manipulated (by adding fences across it to disrupt fluid flow from base to tip), as can the nearby environment (by building a wall curving around the wing tip).

Birch and Dickinson programmed robfly to move its wing in a hovering motion, and the result is the most detailed picture ever obtained of flow over an insect

Aerodynamics

Flight of the robfly

George V. Lauder

Qualitative studies of airflow over insect wings have long been possible, thanks to the use of smoke trails. With a new robotic fly, flow and force can be analysed quantitatively, so theories of insect flight can be tested.

The problem of studying how air moves around flying animals has attracted attention from zoologists, aeronautical engineers and computational fluid dynamicists, but has remained generally unresolved. It is terribly difficult to measure patterns of airflow accurately in three dimensions, especially around insect wings, which are typically small and move rapidly in a complex manner. Yet quantifying such patterns is essential for understanding the aerodynamic mechanisms of insect flight and for testing theories about wing function. On page 729 of this issue, Birch and Dickinson¹ describe how they used a dynamically scaled robotic insect to obtain new data on how insect wings function during hovering. The importance of their work goes beyond the specific hypothesis that they test, and shows the power of a

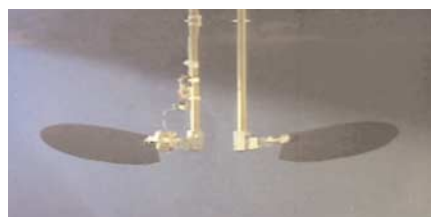


Figure 1 Robofly. Two model fruitfly wings, which can be controlled precisely in three dimensions, are attached to force sensors and immersed in a vat of mineral oil.

laboratory model that combines quantitative analyses of airflow with direct measurements of the forces produced by wings.

Our understanding of the aerodynamics of insect flight has been helped greatly by observations of tethered insects flying in a

wing. A beautiful tip vortex is visible, as is a strong downwash behind the wing, and lateral flow (from the base of the wing towards the tip) is seen along the rear two-thirds of the wing's upper surface (Fig. 2). The leading-edge vortex is also clearly present. However, it does not have the helical structure of the hawkmoth vortex, and fluid within the vortex does not flow significantly from the wing's base to its tip. This finding is noteworthy: leading-edge vortices on flapping wings are unstable and tend to break away, causing a rapid loss of lift. Visualization of smoke trails over the hawkmoth wing⁶ suggested that leading-edge vortices are stabilized by strong lateral helical flow, but this is not apparent on the robofly wing.

So how might robofly stabilize these vortices? To investigate the problem, Birch and Dickinson eliminated all lateral flow by attaching teardrop-shaped fences perpendicular to the wing surface, with the fattest portion of the teardrop at the leading edge. Such fences should block any lateral flow and, if the present view of leading-edge vortices⁶ is correct, should result in decreased lift. But the opposite occurred: the lift forces actually increased slightly when the fences were present. This makes it unlikely that insect equivalents of robofly — fruitflies — stabilize leading-edge vortices by lateral helical flow, and suggests that these vortices could actually grow larger before becoming unstable. Although there is considerable variability, the bodies of most insects are 2 to 4 millimetres long, equivalent to fruitflies. We must seek other mechanisms of vortex stabilization for such insects.

In the future, by changing the viscosity of the mineral-oil bath and the shape of the wings, it should be possible to use robofly to reveal the flow and force patterns around insects with longer wings, 2 to 5 centimetres long. Modification of robofly and the flapper to a state equivalent to forward flight would also be valuable. Ultimately, however, it may be possible to integrate DPIV with micro-electromechanical-systems technology, allowing simultaneous measurements of flow and force around freely flying insects. Then the insects themselves can tell us if our models are correct. ■

George V. Lauder is in the Department of Organic and Evolutionary Biology, and the Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA. e-mail: glauder@oeb.harvard.edu

1. Birch, J. M. & Dickinson, M. H. *Nature* **412**, 729–733 (2001).
2. Grodnitsky, D. L. & Morozov, P. P. *J. Exp. Biol.* **169**, 143–163 (1992).
3. Brodsky, A. K. *J. Exp. Biol.* **161**, 77–95 (1991).
4. Ellington, C. P. *Phil. Trans. R. Soc. Lond. B* **305**, 41–78 (1984).
5. Ellington, C. P. *Phil. Trans. R. Soc. Lond. B* **305**, 79–113 (1984).
6. Ellington, C., van den Berg, C., Willmott, A. P. & Thomas, A. *Nature* **384**, 626–630 (1996).
7. Lauder, G. V. *Am. Zool.* **40**, 101–122 (2000).
8. Dickinson, M. H., Lehmann, F.-O. & Sane, S. *Science* **284**, 1954–1960 (1999).
9. Drucker, E. G. & Lauder, G. V. *J. Exp. Biol.* **202**, 2393–2412 (1999).

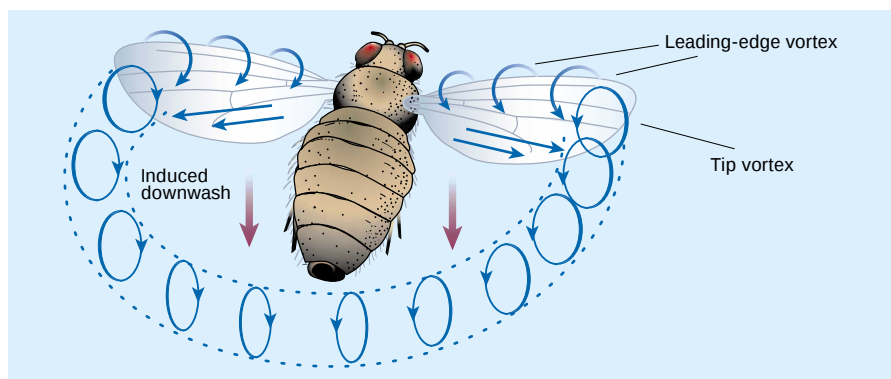


Figure 2 Patterns of airflow during the downstroke of a hovering insect, as revealed by Birch and Dickinson's quantitative analysis¹ of fluid flow over the wing of robofly. Airflow over the leading edge of the wing rolls up into a leading-edge vortex (LEV). LEVs contain a low-pressure core that enhances lift, but they are unstable and tend to detach, causing a rapid reduction in lift. This has prompted research into how insects stabilize LEVs during flight. One proposed mechanism is the lateral flow of air from the base to the tip of the wing⁶. However, when Birch and Dickinson installed barriers to lateral fluid movement on the robofly wing, an increase in lift occurred, suggesting that, for insects of fruitfly size, mechanisms other than lateral air movement must stabilize LEVs. The presence of a strong vortex at the wing tip and downwash behind the wing may stabilize LEVs by reducing the effective angle of wing attack (the angle between the wing and the oncoming air).

Quantum physics

Cooperation includes all atoms

Juha Javanainen

Atom statistics is a fundamental property of all particles that dictates how they behave in certain situations. But behaviour previously attributed to atom statistics is equally likely to arise from cooperative effects.

All microscopic particles can be classified either as bosons (particles such as photons that obey Bose–Einstein statistics) or fermions (particles such as electrons that obey Fermi–Dirac statistics). Although it often takes an elaborate low-temperature experiment to see the difference between bosons and fermions in an atomic system, the distinction is woven into the fabric of the Universe. If electrons were bosons instead of fermions, matter would be different in unimaginable ways, and we would probably not be around to contemplate the issue. For example, the fermionic nature of electrons is responsible for the chemistry of the periodic table. Nonetheless, as two groups from Arizona¹ and MIT² argue in *Physical Review Letters*, behaviour that physicists have attributed to atoms being bosons may actually arise from cooperative dynamics between the atoms, and therefore would be equally valid for fermionic atoms.

The case in point is four-wave mixing, a standard tool of nonlinear optics in which two light waves interfere to form a periodic pattern. A third wave diffracted from this pattern generates a fourth wave at a particular frequency. The 1995 achievement of Bose–Einstein condensation — a form of matter in which all the atoms are in the same quantum state — has allowed the development of

analogous experiments with atoms³. In this case, there are two beams of atoms running at each other. By virtue of quantum mechanics, there is a wave associated with the moving atoms. So the two atomic waves interfere and make a standing-wave pattern — a grating of atoms. A 'probe' atom coming along will act as a wave in its own right and diffract off the grating, resulting in a fourth atomic wave travelling in a particular direction.

The results of experiments on four-wave mixing are usually interpreted as a consequence of boson statistics, specifically the enhancement or amplification of the bosonic atoms or photons. Such behaviour is forbidden by fermion statistics, so other bosonic processes, such as amplification of light or matter waves, would also appear to be ruled out for fermions. But the Arizona¹ and MIT² groups show that these processes can indeed occur with fermions. Happily, no violations of atom statistics are required, simply a more mature understanding of how quantum mechanics works.

Consider a more quantitative picture of four-wave mixing with atoms, in which the first two beams, 1 and 2, each have N bosonic atoms with momenta $\mathbf{k}_1$ and $\mathbf{k}_2$. Atom–atom interactions cause an atom to transfer from beam 1 to beam 2, and the conservation of momentum means that the probe atom has

to deflect in a new direction, generating the fourth wave. The probability that this process occurs is proportional to the number of atoms, N , available to scatter in beam 1. But Bose–Einstein statistics says that bosons prefer to join already existing bosons, so the probability is also proportional to the number of atoms, N , in beam 2 that receives the probe atom. The resulting N^2 dependence of the probability for the probe atom to scatter in a certain direction can therefore be attributed to bosonic enhancement.

This, though, is not the only conceivable description of four-wave mixing^{1,2}. According to quantum mechanics, to find a probability for a transition between two states of a system one first finds a transition amplitude, f , and then takes the square of it. In four-wave mixing there are N transition amplitudes for the scattering of the probe atom from the initial state to the final state, one for each atom that could transfer between the two beams. Another intriguing provision of quantum mechanics states that if in a given experiment it is impossible, even in principle, to distinguish between the paths that lead from the same initial state to the same final state, then the transition amplitudes have to be added. If this holds, then the probability amplitude for four-wave mixing becomes Nf and the probability again scales as N^2 . This time, though, the reason is not bosonic enhancement, but another common cause for N^2 dependence in physics: cooperation. Although only one atom was transferred between beams 1 and 2, all the atoms acted in concert to boost the transition amplitude.

The role of atom statistics is therefore only secondary. To date, all the experiments on four-wave mixing³ and analogous amplification of matter waves^{4,5} have been done with bosons because they conveniently provide atom waves that have sharply defined wavelengths. But one might imagine doing the same thing with fermionic atoms. In a Bose–Einstein condensate there is no way to distinguish between the atoms, and cooperation is assured. The defining property of fermions is that only one fermion can occupy one quantum state, so N fermionic atoms cannot be prepared with the same momentum, $\mathbf{k}_1$ or $\mathbf{k}_2$. Still, it is feasible to have atoms with momenta in a narrow enough range around $\mathbf{k}_1$ and $\mathbf{k}_2$ so that the spread cannot be resolved experimentally. Then it is not possible to distinguish between the transition paths, cooperation holds, and the N^2 dependence on the number of atoms emerges. This time, though, the atoms are fermions, and bosonic enhancement is not a viable explanation.

As a classic example of cooperation, imagine a system of N excited atoms. An atom is coupled to the ever-present electromagnetic fields, so an excited atom will spontaneously emit a photon. But if the N atoms reside at equivalent positions relative to the field, they all couple to the electromagnetic fields in the

same way and cannot be distinguished by the way they interact with the field. Spontaneous emission from the atoms is then cooperative⁶: the N excited atoms return to their ground state by spontaneously emitting a flash of light whose peak intensity scales as N^2 . This dramatic process is called superradiance, but it is a subtle matter of interpretation whether it has been seen experimentally. Superradiance as described by Dicke⁶ is one of the key paradigms in optical physics, but an unequivocal demonstration would need the atoms to be placed in a region smaller than the wavelength of the light to fulfill the requirement for 'equivalent field positions'. Even if this were practicable, the interactions between the atoms would spoil cooperation anyway. The moral of this story is that experimental realities tend to obstruct cooperative behaviour.

In my opinion, the main contribution of the Arizona¹ and MIT² work is that it demonstrates the fragility of the distinction between atom statistics and cooperative behaviour. Physicists now know that cooperative behaviour may mock Bose–Einstein statistics. Perhaps the converse is true: can Bose–Einstein and Fermi–Dirac statistics be harnessed to assist in cooperative phenomena, such as superradiance?

Juha Javanainen is in the Department of Physics, University of Connecticut, Storrs, Connecticut 06269-3046, USA.

e-mail: jj@phys.uconn.edu

1. Moore, M. G. & Meystre, P. *Phys. Rev. Lett.* **86**, 4199–4202 (2001).
2. Ketterle, W. & Inouye, S. *Phys. Rev. Lett.* **86**, 4203–4206 (2001).
3. Deng, L. *et al. Nature* **398**, 218–220 (1999).
4. Inouye, S. *et al. Nature* **402**, 641–644 (1999).
5. Kozuma, M. *et al. Science* **286**, 2309–2312 (1999).
6. Dicke, R. H. *Phys. Rev.* **93**, 99–110 (1954).

Neurobiology

Stem cells on the brain

Robert Cassidy and Jonas Frisén

Stem cells have great potential for treating a variety of diseases and seem to hit the headlines almost every week. An extremely pure population of brain stem cells has now been obtained from adult mice.

Many of us were taught in school that we are born with a certain number of nerve cells, which cannot renew themselves. But, after some initial indications in the 1960s, it is now known that our brains are much more capable of change than was previously thought, and that neurons are continuously replenished in some regions of the adult brain¹. These new neurons are derived from immature cells called stem cells. Stem cells have the potential to

generate not only new stem cells but also several types of mature cell; neural stem cells, for example, can produce both neurons and supporting cells called glia. This potential has fuelled hope that brain stem cells might be used in the treatment of, for instance, neurodegenerative disorders such as Parkinson's disease². Yet first we must be able to isolate these cells, and this in turn requires a knowledge of their defining features. On page 736 of this issue³, Rietze and colleagues

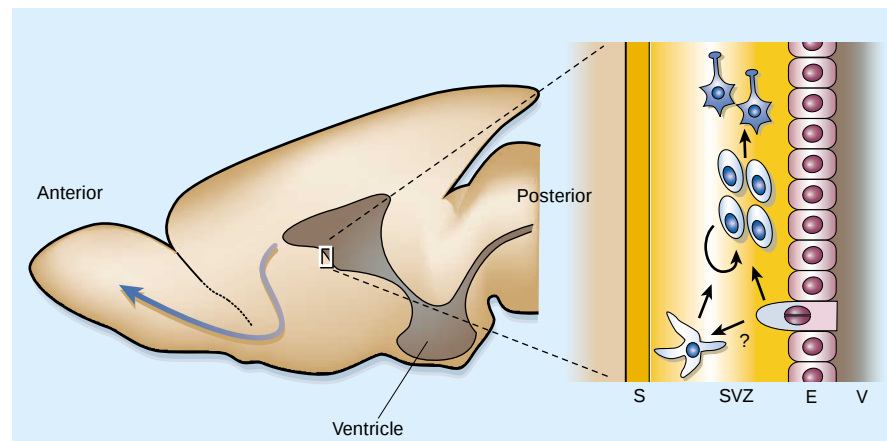


Figure 1 A source of brain stem cells. Left, a section of the mouse brain (as if it were cut down the midline, from top to bottom), showing the ventricle wall. Inset, details of the ventricle wall. In adult mammalian brains, neurons are born in the subventricular zone (SVZ) between the ependymal layer (E) and the striatum (S) of the anterior lateral ventricle (V), and migrate along the rostral migratory stream to the olfactory bulb (indicated by the long blue arrow in the left-hand part). In this model, stem cells in the ependymal layer generate a population of rapidly dividing 'progenitor' cells (ovals). These in turn generate the newborn neurons (dark cells). Astrocytes (branched cell) could be independent stem cells, or intermediate cells between ependymal stem cells and the progenitor cells.

tackle the problem, describing how they achieved a high degree of purification of neural stem cells.

Within the brain, stem cells reside in the walls of ventricles — cavities that are filled with cerebrospinal fluid. In this region, only about 1 in 300 cells (0.3%) is a stem cell³, so they are incredibly difficult to isolate. Indeed, although neural stem cells in the developing peripheral nervous system (outside the brain and spinal cord) have been identified and purified⁴, brain stem cells have never before been isolated at more than 5% purity, with different antibodies being used to identify embryonic⁵ and adult⁶ cells. Rietze *et al.*³ have now isolated adult brain stem cells at an impressive 80% purity.

First, the authors³ fractionated brain cells from the lateral ventricle wall of mice (Fig. 1), sorting cells according to their size and the molecules expressed on their surface. They then investigated the properties of cells of different sizes and combinations of cell-surface proteins *in vitro*. One population of cells proved particularly interesting. These had a diameter larger than 12 μm . They expressed little or no heat-stable antigen, one type of cell-surface marker, and showed none or only weak binding to peanut agglutinin — a marker previously used to isolate mouse haematopoietic stem cells⁷. The authors found that 80% of this population of brain cells had properties of stem cells *in vitro*.

Moreover, when these purified neural stem cells were transplanted to the developing brains of mouse embryos, they produced neurons and glia, which appeared to integrate into the host brains. An indication that they may act as neural stem cells under physiological conditions came from the authors' analysis³ of a mutant mouse dubbed *querkopf*⁸. In these mice, the production of new neurons in the lateral ventricle wall is much reduced⁸, as is the number of cells with the characteristics described by Rietze *et al.*³.

Over the past few years, other types of stem cell have proved very versatile. If these cells are transferred to a new environment, they can generate mature cells that are different to those they would normally produce. This implies that the differentiation of their progeny is not only intrinsically regulated, but is also influenced by the environment. Neural stem cells can also generate a large variety of non-neural cell types under experimental conditions, and several studies have shown that they can produce muscle cells^{9–11}. Rietze *et al.* tested whether the cell type they isolated also has this potential, and found that more than 50% of these cells generated muscle cells when plated on a muscle-cell line. The authors were able to follow single purified brain cells as they differentiated to become muscle cells. Although this phenomenon of stem-cell plasticity is well established, little is known about its molecular basis. The efficient differentiation

of single purified neural stem cells to muscle cells *in vitro* provides a valuable model for molecular analyses.

How do Rietze *et al.*'s cells relate to previously described neural stem cells? The ventricles of the central nervous system are lined by a single layer of cells, called ependymal cells. In certain regions there is a narrow zone of tissue, the subventricular zone, beneath the ependymal layer. Two types of cell have been shown to be neural stem cells, namely ependymal cells⁶ and astrocytes that reside in the subventricular zone¹². Moreover, under certain *in vitro* conditions, other brain cells also display characteristics of stem cells^{13,14}. Rietze *et al.* found that about one-third of the stem cells they isolated reside in the ependymal layer, and two-thirds in the subventricular zone.

So are these purified stem cells subventricular astrocytes or ependymal cells? They do not have extensions known as cilia, a characteristic of ependymal cells, and they do not express glial fibrillary acidic protein, an astrocyte-specific protein. But this does not in itself answer the question — the purification techniques used often strip cilia from ependymal cells, and the absence of an astrocyte-specific protein may be a result of inefficient detection. So it remains unclear how the purified cells relate to previously described neural stem cells. Moreover, because the

authors used a negative-selection technique (that is, the absence of heat-stable antigen and lack of binding to peanut agglutinin), there is still no specific positive marker for neural stem cells. Yet the ability to purify brain stem cells to this degree will be invaluable for future studies, and may prove useful in developing treatments. ■

Robert Cassidy and Jonas Frisén are in the Department of Cell and Molecular Biology, Medical Nobel Institute, Karolinska Institute, SE-171 77 Stockholm, Sweden.

e-mails: rob.cassidy@cmb.ki.se

jonas.frisen@cmb.ki.se

1. Gross, C. G. *Nature Rev. Neurosci.* **1**, 67–73 (2000).
2. Björklund, A. & Lindvall, O. *Nature Neurosci.* **3**, 537–544 (2000).
3. Rietze, R. L., Valcanis, H. & Bartlett, P. F. *Nature* **412**, 736–739 (2001).
4. Morrison, S. J., White, P. M., Zock, C. & Anderson, D. J. *Cell* **96**, 737–749 (1999).
5. Uchida, N. *et al. Proc. Natl Acad. Sci. USA* **97**, 14720–14725 (2000).
6. Johansson, C. B. *et al. Cell* **96**, 25–34 (1999).
7. Reisner, Y., Itzicovitch, L., Meshorer, A. & Sharon, N. *Proc. Natl Acad. Sci. USA* **75**, 2933–2936 (1978).
8. Thomas, T., Voss, A. K., Chowdhury, K. & Gruss, P. *Development* **127**, 2537–2548 (2000).
9. Clarke, D. L. *et al. Science* **288**, 1660–1663 (2000).
10. Galli, R. *et al. Nature Neurosci.* **3**, 986–991 (2000).
11. Tsai, R. Y. L. & McKay, R. D. G. *J. Neurosci.* **20**, 3725–3735 (2000).
12. Doetsch, F., Caille, I., Lim, D. A., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. *Cell* **97**, 703–716 (1999).
13. Palmer, T. D., Markakis, E. A., Willhoite, A. R., Safar, F. & Gage, F. H. *J. Neurosci.* **19**, 8487–8497 (1999).
14. Kondo, T. & Raff, M. *Science* **289**, 1754–1757 (2000).

Earth science

Gas hydrates and deglaciations

Stein B. Jacobsen

Long ago, Earth experienced a series of especially severe glaciations. A new explanation for deposits laid down at the end of these events centres on the large-scale release of methane from solid gas hydrates.

The Earth's harshest glaciations occurred between about 600 million and 800 million years ago, in the late Neoproterozoic era. The severity of the glaciations was first recognized¹ in the 1960s. Some 30 years later came the suggestion that during these intervals the entire Earth was covered with ice, hence the name by which this idea has become known² — snowball Earth.

There is now strong evidence for at least four severe glaciations during the Neoproterozoic^{3,4}. One explanation⁵ for how they end centres on the warming effect of atmospheric CO₂, with the subsequent fate of the CO₂ being evident in a special type of carbonate rock called cap carbonates, which overlie late Neoproterozoic glacial sediments on almost every continent. In contrast, in a paper in *Geology*, Kennedy *et al.*⁶ invoke massive methane release as a consequence of the deglaciation and call into question some of the thinking about a snowball Earth. Kennedy *et al.* propose that the cap carbon-

ates were deposited following the release of methane from gas hydrates in terrestrial permafrost, after rapid warming at the end of the glaciations and flooding of widely exposed continental shelves and terrestrial basins. The flooding is explained by melting of the glacial ice, with warming of the flooded permafrost leading to methane release through destabilization of gas hydrates. Gas hydrate is a solid, ice-like substance composed of rigid cages of water molecules that enclose molecules of methane. In their model, Kennedy *et al.* do not comment on how the warming is thought to have been initiated.

The most comprehensive snowball Earth explanation for the late Neoproterozoic glaciations is that of Hoffman *et al.*⁵, whose argument runs as follows. During that time, large parts of the continental landmasses were at middle to low latitudes, a situation that has not occurred since then. This configuration could have produced global glacial conditions because continents are good



100 YEARS AGO

Wireless telegraphy is in use upon ships engaged in the naval manoeuvres, and it enables a battleship to communicate with a cruiser fifty or sixty miles away with greater ease than the same ship could be communicated with at a distance of ten miles in clear weather a year or two ago. But the *Times* special correspondent with one of the fleets remarks that the method, although independent of the weather, is still subject to one very serious drawback. The communication is not, and cannot be, a private or exclusive one, except so far as the messages are transmitted in cipher... The moral is to employ as difficult a cipher as possible and to change it as soon as there is any reason to suspect the enemy has discovered it... It will never be quite satisfactory for war purposes until the transmitting instrument can be so adjusted as to emit vibrations of different pitch at the will of the operator and the receiving instrument rendered sensitive only to vibrations of a given pitch at a given moment.
From *Nature* 15 August 1901.

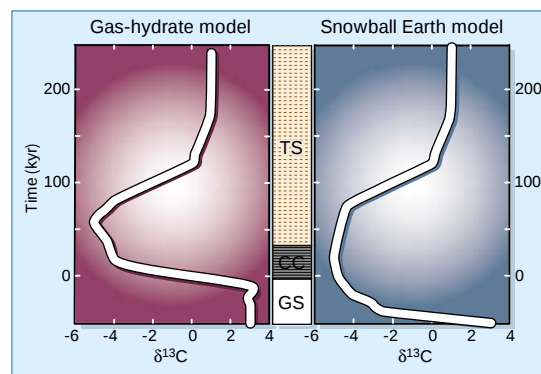
50 YEARS AGO

A Naturalist in the Gran Chaco. There is a rich tradition of naturalist travellers in the University of Edinburgh... and this record of Sir John Graham Kerr's work in the Gran Chaco of South America places him worthily in their distinguished company... The whole story hangs by a thread — on a bleak afternoon in February 1889 an Edinburgh medical student picks up a copy of *Nature* at the bookstall in Waverley Station, sees a note about a projected Argentine expedition to certain high tributaries of the Rio de la Plata, and so the adventure is conceived and Graham Kerr's scientific future potentially fashioned. As naturalist on Captain Juan Page's expedition, the author penetrated the Gran Chaco, and his observations, made during 1889–91, form the greater part of the book. The second part describes the Lepidosiren Expedition in the same region, undertaken with John Budgett during 1896–97. The first journey... was a naturalist's roving commission, and every observation in this unobserved country came as grist to the young and enthusiastic traveller. The final and mysterious disappearance of his collections after he had left the river-steamer *Bolivia* was a serious but not a crippling loss, for the vital contributions to knowledge were the field observations recorded in this book.
From *Nature* 18 August 1951.

Box 1 Carbon isotopes and model expectations

Carbon has two stable isotopes, with about 99% of carbon in carbonate rocks being ^{12}C and the rest ^{13}C . The isotopic composition of carbon is generally shown in the $\delta^{13}\text{C}$ notation that represents the deviation of the $^{13}\text{C}/^{12}\text{C}$ ratio in parts per thousand (‰) from a laboratory standard. Photosynthetic organisms preferentially incorporate $^{12}\text{CO}_2$ relative to $^{13}\text{CO}_2$. This process results in marine organic matter being depleted in ^{13}C ($\delta^{13}\text{C} \sim -25\text{‰}$) relative to seawater bicarbonate ions ($\delta^{13}\text{C} \sim 0$ to $+5\text{‰}$).

The isotopic composition of mantle-derived carbon is known ($\delta^{13}\text{C} = -6 \pm 1\text{‰}$). So a mass-balance calculation can show that, over the past 3.5 billion years, the amount of carbon buried in sediments in the Earth's crust as dead organic matter was about 30% compared with about 70% buried as inorganic carbon in



carbonate rocks. The gas-hydrate and snowball Earth models offer different predictions of $\delta^{13}\text{C}$ perturbation towards the end of, and following, a glaciation, as shown in the figure here. The timescale is from Kennedy *et al.*⁶, with time zero set at the end of the glacial interval and initiation of cap-carbonate deposition. GS, glacial sediments; CC, cap carbonate; TS, turbidite sediments.

The stratigraphic relationships support the idea that the negative carbon isotopic shift developed abruptly during deglaciation, as the gas-hydrate model⁶ predicts, and that the recovery to more normal isotopic values required about 100,000 years. The total duration of the glacial interval in the snowball Earth model needs to be at least ten million years, compared with 0.1–1 million years for the gas-hydrate model. S.B.J.

reflectors of solar energy (today, by contrast, most of the solar energy absorbed by the Earth is trapped in the tropical oceans). According to this model⁵, a continuous lid of sea ice, separating the CO_2 reservoirs in the ocean and the atmosphere, would have covered the oceans. Escape from snowball conditions would have been due to the gradual build-up in the atmosphere of the greenhouse gas CO_2 , emitted from volcanoes. In Hoffman and colleagues' model, diminished weathering of continental silicates during the glaciation meant that levels of CO_2 reached 120,000 parts per million (today's value is about 366 parts per million), triggering rapid deglaciation and melting of sea ice. Weathering then drew down CO_2 , producing ocean alkalinity and precipitating the cap carbonate.

This is the 'hard' version of snowball Earth, one in which every part of the oceans was covered in ice. Many climate modellers have problems with it^{7,8}, not the least being how a frozen tropical ocean could be maintained with extremely high levels of CO_2 in the atmosphere⁷. Moreover, in the long term Earth's surface temperature has been remarkably stable, which is thought to be due to a feedback mechanism between temperature, weathering rate and the concentrations of atmospheric CO_2 (ref. 9). Such long-term models do not include any snowball states for the Earth. And more recent attempts to model 'hard' snowball glacia-

tions have failed, from which it has been concluded that ice-covered continents, but with an equatorial belt of open oceans, are more consistent with the geological evidence^{7,8}.

The carbon-isotope variation ($\delta^{13}\text{C}$) seen in cap carbonates is central to explanations for the end of the late Neoproterozoic glaciations. The details are given in Box 1. According to the snowball Earth model, during a glaciation there would be a short-lived change in the carbon isotopic composition of the entire ocean because of the essentially complete elimination of the marine biosphere. The result would be a rapid decrease of $\delta^{13}\text{C}$ in marine carbonate rocks to $-6 \pm 1\text{‰}$. Recovery to normal values ($> 0\text{‰}$) would occur in postglacial times. In contrast, in the gas-hydrate model, $\delta^{13}\text{C}$ during glacial times would remain unchanged, followed by a sharp postglacial fall in its value, owing to destabilization of methane gas hydrates (which have a $\delta^{13}\text{C}$ of around -60‰). In this model, most of the methane released from the sediments was oxidized at the sediment–seawater interface by sulphate reduction, producing alkalinity and carbonate precipitation.

Observed carbon isotope values³ in cap carbonates globally range from ~ 0 to -5‰ , but the isotopic pattern is well defined in only a few places. One — as Kennedy *et al.*⁶ point out — is the southern margin of an ancient rock structure in Africa, the Congo

craton, where the pattern matches that expected from the gas-hydrate model. Further, values of strontium isotopes during the glacial intervals provide no support for greatly diminished weathering during glaciation, or for greatly enhanced weathering in its immediate aftermath⁴, as required by the snowball Earth model. But they are consistent with the gas-hydrate hypothesis.

Both hypotheses involve short-lived changes in the carbon-isotopic composition of the ocean. But the gas-hydrate model avoids some of the obvious difficulties associated with huge changes in atmospheric CO₂ in the snowball Earth model, such as the long time (ten million years) required for such large concentrations to build up. It also provides a better explanation of the temporal variation of carbon isotopes in marine carbonates through the late Neoproterozoic.

But although Kennedy *et al.* have put forward a powerful case, this is, no doubt, far from the end of the story. ■

Stein B. Jacobsen is in the Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, Massachusetts 02138, USA. e-mail: jacobsen@neodymium.harvard.edu

1. Harland, W. B. *Geol. Rundsch.* **54**, 45–61 (1964).
2. Kirshvink, J. L. in *The Proterozoic Biosphere: A Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 51–52 (Cambridge Univ. Press, New York, 1992).
3. Kaufman, A. J., Knoll, G. H. & Narbonne, G. M. *Proc. Natl Acad. Sci. USA* **94**, 6600–6605 (1997).
4. Jacobsen, S. B. & Kaufman, A. J. *Chem. Geol.* **161**, 37–57 (1999).
5. Hoffman, P. F. *et al. Science* **281**, 1342–1346 (1998).
6. Kennedy, M. J., Christie-Blick, N. & Sohl, L. E. *Geology* **29**, 443–446 (2001).
7. Hyde, W. T., Crowley, T. J., Baum, S. K. & Peltier, W. R. *Nature* **405**, 425–429 (2000).
8. Crowley, T. J., Hyde, W. T. & Peltier, W. R. *Geophys. Res. Lett.* **28**, 283–286 (2001).
9. Walker, J. C. G., Hays, P. B. & Kasting, J. F. *J. Geophys. Res.* **86**, 9776–9782 (1981).

Neuroscience

Dynamic categories

Michael P. Kilgard

Neuroscientists often study brain function by presenting a stimulus many times and averaging the neural response. A trick for finding brain activity in single trials might reveal how animals mentally categorize information.

One of the most enduring problems in biology is that of how the brain assigns meaning to the stimuli it receives. We naturally divide the world into convenient categories, but the brain mechanisms that underlie this ability are poorly understood. On page 733 of this issue, Ohl and colleagues¹ report that category formation in animals is accompanied by fleeting but identifiable patterns of brain activity. The patterns exist for only a fraction of a second, and would be dismissed as noise by more traditional methods of analysis. Ohl *et al.*'s powerful technique may eventually reveal whether and how similar patterns contribute to our own cognitive abilities.

Historically, the responses of human and animal brains to a stimulus, such as a particular sight or sound, have been analysed by presenting the stimulus many times and averaging the neuronal activity that occurs. But because the responses are averaged, this approach detects only activity that occurs reliably at a certain time after every stimulus (such activity is said to be 'time-locked' to the stimulus). Nonetheless, the technique has yielded considerable insight into brain function.

One type of response, called mismatch negativity (MMN), has been widely used to study how the brain categorizes similar stimuli. This type of brain activity occurs when an unexpected sound (a mismatch) interrupts a series of identical auditory stimuli. Categorization is particularly important in

processing spoken language, where a variety of sounds are given one meaning. For example, the gender, age and emotion of the speaker significantly affect the physical qualities of speech, but 'dog' means 'dog' no matter who says it or in what tone. Of course, speech sounds may have varying meanings in different languages. The MMN response has been used to show that native and non-native speakers of a particular language categorize sounds from that language differently^{2,3}. Language-specific representations of speech sounds develop in infants between six and twelve months of age⁴, but can be altered throughout adulthood with sufficient practice⁵.

Although these studies provide a physiological marker of categorization that changes with learning, the MMN response itself does not represent specific categories, and the neural basis for categorization has remained elusive. To tackle the problem, Ohl and colleagues previously trained gerbils to categorize tones as either rising or falling in pitch⁶. After a few days of training, the animals could categorize a new tone instantly, without a need for training with that stimulus. On hearing a rising tone they moved immediately to another compartment of the box in which they were housed; after hearing a falling tone, they stayed where they were. Lesion of various parts of the brain showed that the right auditory cortex is required to discriminate between rising and falling tones⁷. Neurons in this region respond dif-

ferently to tones that are sweeping upwards and downwards. However, these differences seem to reflect the physical differences in the sounds, rather than their behavioural meaning (their category).

Largely on the basis of studies of the olfactory system, Freeman and colleagues suggested that dynamic patterns of brain activity carry information about the meaning of a stimulus⁸. But these patterns are not reliably time-locked to the stimulus, so they cannot be detected in averages of neuronal activity. Ohl, Scheich and Freeman¹ have now developed a strategy for identifying unusual activity patterns in a single trial, and applied it to the auditory system of gerbils. They find that certain activity patterns, which they call marked states, occur in animals that have learned to distinguish between rising and falling tones (Fig. 1, overleaf). The authors suggest that these states might provide a mechanism by which the animals can recognize the abstract qualities that define a category.

Ohl *et al.*¹ detected the unusual brain states by using a grid of 18 electrodes, positioned directly over the auditory cortex, that detect electroencephalographic (EEG) activity. To isolate 'local' activity patterns that occur within the auditory cortex, the authors removed any activity that was common to all electrodes. The patterns of activity that occurred in response to a single rising tone were then compared with the averaged pattern from 30 trials with falling tones, and vice versa. Unusual patterns were considered marked states if they were more than three standard deviations away from the average pattern produced in response to the opposite tone.

As expected⁹, the rising and falling tones evoked different patterns of activity in the auditory cortex in trained and untrained animals¹. Then, as each animal learned to distinguish the sounds, further marked states developed within a few seconds of each tone¹ (Fig. 1c, black boxes). The crucial finding is that these dynamic patterns were stabilized when animals learned that the behavioural meaning of a stimulus ('move' or 'don't move') was based on the abstract quality of tone direction. Although this finding was replicated in only four animals, the discovery of these activity patterns is an important step towards identifying the biological basis of categorization. It appears that the key to the authors' success lay in removing the irrelevant (global) activity to reveal local patterns that were most closely related to the behavioural meaning of the stimuli. Future work will doubtless look at how the patterns vary from trial to trial and from individual to individual, and whether marked states are related in any way to the neuronal patterns evoked by the two classes of stimulus¹⁰.

The marked states developed during the decision-making period, before each animal was forced to choose whether to move to another compartment. It will be interesting

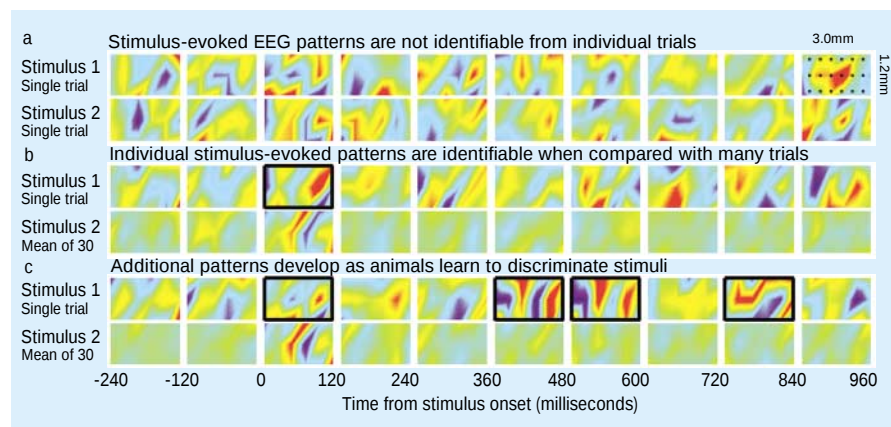


Figure 1 Detecting the brain activity that underlies the formation of categories. Each rectangle represents the pattern of activity in the auditory cortex at an instant in time in response to one of two sounds. Red denotes high activity. The black dots mark the locations of the electrodes recording neuronal activity. **a**, Brain activity can be very noisy, making it hard to identify patterns. **b**, Ohl *et al.*¹ have developed a technique for identifying the unique patterns that arise at some point after each of the sounds is heard. The strategy involves averaging the responses to stimulus 2 (or 1), and comparing them with individual responses to stimulus 1 (or 2). **c**, Additional patterns develop when animals learn to categorize the stimuli. Black boxes mark the patterns of brain activity that can be identified in a single trial using this technique.

to find out how the timing of these states changes when the task contingencies (such as the length of the decision-making period) are altered, and whether similar states develop during other categorizations. Ohl *et al.*¹ have already shown that marked states do not develop in the visual cortex during this auditory task. We now need to know whether they develop in the visual cortex during the categorization of visual stimuli.

Can this new technique be used to study how humans form categories? The number of EEG electrodes used in human studies has been increasing rapidly, yielding higher spatial resolution. So far, most groups have analysed only averaged brain responses. More sophisticated analysis of high-resolution EEG data will probably reveal complex local patterns in humans. The greater challenge will be in making sense of these patterns. Ohl

et al.'s technique may not lead to mind reading, but it does suggest a variety of new tools that may yet reveal the cognitive processes that give meaning to our world. ■

Michael P. Kilgard is at the School of Human Development, University of Texas at Dallas, Richardson, Texas 75083, USA.
e-mail: kilgard@utdallas.edu

1. Ohl, F. W., Scheich, H. & Freeman, W. J. *Nature* **412**, 733–736 (2001).
2. Näätänen, R. *et al.* *Nature* **385**, 432–434 (1997).
3. Sharma, A. & Dorman, M. F. *J. Acoust. Soc. Am.* **107**, 2697–2703 (2000).
4. Cheour, M. *et al.* *Nature Neurosci.* **1**, 351–353 (1998).
5. Tremblay, K., Kraus, N., Carrell, T. D. & McGee, T. J. *Acoust. Soc. Am.* **102**, 3762–3773 (1997).
6. Wetzels, W., Wagner, T., Ohl, F. W. & Scheich, H. *Behav. Brain Res.* **91**, 29–39 (1998).
7. Wetzels, W., Ohl, F. W., Wagner, T. & Scheich, H. *Neurosci. Lett.* **252**, 115–118 (1998).
8. Freeman, W. J. *Prog. Brain Res.* **102**, 319–333 (1994).
9. Ohl, F. W., Schulze, H., Scheich, H. & Freeman, W. J. *J. Physiol. (Paris)* **94**, 549–554 (2000).

Planetary science

A new model Moon

Jay Melosh

The most sophisticated simulations yet of the Moon's birth show that it could have been created from an impact of a large body with a fully formed, rather than half-formed, Earth.

The idea that the Moon is a by-product of a stupendous collision between the early Earth and a Mars-size protoplanet has been considered respectable science for more than a decade. It is now the consensus view¹, for no other theory reconciles so many facts about the Earth and Moon (Fig. 1) and provides a relatively seamless explanation of how our large satellite came to be. But there have

been few attempts to directly model the complex physical processes occurring during a planetary impact. Cameron² was one of the first people to suggest the giant impact theory, Hartmann³ being the other principal proponent of the idea in the early days. Cameron and his collaborators have since led the efforts to model the impact numerically, and their view of events has been highly influential.

Now, however, thanks to the advance of computer technology and the development of better impact algorithms, other groups are also able to simulate collisions between planets. On page 708 of this issue⁴, Canup and Asphaug usher in this new era by describing the outcome of the giant impact in terms that in many respects differ from Cameron's latest results⁵.

Computer simulation of a planetary-scale impact is not a task for faint-hearted, point-and-click computer modellers. Not only does such a collision involve all the details of shock physics, melting and vaporization, but the mutual gravitational interactions of all those hot fluids squirting around in space also have to be taken into account. Furthermore, to impart the angular momentum that keeps the Moon in orbit, the impact must have been glancing, not head on, so the problem must be treated in three dimensions. It has taken squadrons of physicists in the United States, Russia and elsewhere nearly 50 years to come up with computers and three-dimensional computer codes that can adequately treat the effects of impacts and explosions under relatively simple conditions in which self-gravity is not important. Adding self-gravity to these codes therefore posed a formidable challenge.

A solution recently arose from an approach⁶ called SPH (smooth particle hydrodynamics), which can, in principle, treat both shock physics and gravity. Benz and Cameron were the first⁷ to apply this method to the origin of the Moon. Other early models^{8,9} relied on conventional 'hydrocodes' that used central gravity, but not self-gravity, and so could accurately model only the early phases of the impact.

The SPH method involves splitting the proto-Earth and the object it collided with into many small computational lumps called 'particles', then computing the interactions between these entities. The results of this procedure become more accurate as more particles are used. Because of computer limitations, the early simulations used only a few thousand particles. It was several years before users recognized that the SPH method requires many more of these particles than originally anticipated. One of the reasons that Canup and Asphaug⁴, who also use an SPH code, achieved results that differ from those of Cameron *et al.* is that they use at least 20,000 particles and so resolve the shock waves and other collision phenomena with much higher accuracy.

Based on his latest simulations⁵, Cameron has held that the collision that created the Moon occurred when the Earth was only about half-formed, about 30 million years after the formation of the first meteorites, as it grew by accretion of material in the early Solar System. This result makes it difficult to understand how the Earth continued to grow without large amounts of

Daedalus

Blood and iron

Blood, says Daedalus, gives a lot of trouble. The heart has the job of pumping it round the body all the time, and can fail or underperform in many ways. Replacement valves and pacemakers help out, as do complete replacement pumps. Sadly, artificial materials tend to encourage the blood to clot. Human hearts from donors are perhaps the best way to overcome this type of heart inadequacy, but have many problems of their own. Daedalus has a new idea.

Blood, he points out, is itself a magnetic fluid. It contains iron-loaded red blood cells, so it could be pumped around the body by a magnetic linear motor. In fact only the red cells really need to be moved, for they provide oxygen. But a fluid loaded with magnetic particles would be pumped as a whole, just as the heart pumps it. The best way of driving the blood around would use coils or magnetic linear motors attached to the arms or legs. Our limbs have central arteries taking blood outwards, and veins near the surface bringing it back in. The veins would be engaged by the motors; the arteries are too deep. One problem will be calibrating the whole system, to establish the correct rate of flow. Daedalus reckons that the carotid arteries in the neck, used as calibrators by the heart itself, could do this job. They will be given calibration coils to check that circulation is maintained. Indeed, the blood pressure could be held at a youthful 120/80 millimetres of mercury, even in very tense moments.

This arrangement has several advantages. No skin need be broken. Even a very inadequate heart can help the arrangement along. Indeed, all patients must have a functioning heart of some sort before inviting surgical help, and it would be a pity not to use it. Even a heart that can only maintain a resting body will still be doing most of the work needed most of the time. The carotid sensors will enable the magnetic accelerators to add just the extra capacity needed from moment to moment. Sadly, the motors on arms and legs will be heavy, even if DREADCO's engineers make the minimal use of copper and iron, and they will need a lot of power. The user will spend a great deal of time changing or recharging batteries. But thousands of customers will be very happy just to have a bit of extra heart-power available when needed. And surgeons will welcome any reduction in the long burden of seeking donors.

David Jones

iron and volatile material also ending up in the iron- and volatile-poor Moon. However, most of the simulations Cameron relied on to constrain the possible range of impacts involved only 3,000 particles. Canup and Asphaug⁴, using a much denser array of particles, now find that an impact on a fully formed Earth, some 50 million to 70 million years after the first meteorites, can loft enough mass to account for the present Moon, thus evading the problems inherent in the early-Moon scheme of events.

Encouraging as these new results⁴ are, they are not the final word. One of the pillars on which any shock-physics computer code rests is the equation of state, a thermodynamic relation between internal energy, density and pressure. Unfortunately this equation is not well known for the complex silicates of which the Earth and Moon are mainly composed. Several difficulties have recently appeared in the most sophisticated equation of state in common use, a computer program called ANEOS¹⁰. Canup and Asphaug⁴ sidestepped these difficulties by using a more primitive equation devised by J. H. Tillotson¹¹. Although it was constructed for impact computations, this equation of state has many drawbacks, including the lack of a clear distinction between solid, melt and vapour phases. So, although vapour pressure

seems to have been involved in boosting vaporized Earth material into orbit, Canup and Asphaug can only evaluate its role by a kind of average over the melt and vapour states. This is an area needing refinements.

Although we have not yet arrived at a definitive simulation of the Moon-forming impact, the signs are encouraging. Thanks to the coevolution of modern computers and simulation algorithms, the means to model events of this kind can now be found on the desk of nearly any scientist. I look forward to the entry of more new players into this fertile scientific arena.

Jay Melosh is in the Lunar and Planetary Lab West, University of Arizona, 1040 East Fourth Street, Gould-Simpson Building, Room 901, Tucson, Arizona 85721-0077, USA.
e-mail: jmelosh@lpl.arizona.edu

1. Canup, R. M. & Righter, K. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) xiii–xv (Univ. Arizona Press, Tucson, 2000).
2. Cameron, A. G. W. & Ward, W. R. *Lunar Sci.* **7**, 120–122 (1976).
3. Hartmann, W. K. & Davis, D. R. *Icarus* **24**, 504–515 (1975).
4. Canup, R. M. & Asphaug, E. *Nature* **412**, 708–712 (2001).
5. Cameron, A. G. W. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 133–144 (Univ. Arizona Press, Tucson, 2000).
6. Monaghan, J. J. *Annu. Rev. Astron. Astrophys.* **30**, 543–574 (1992).
7. Benz, W., Slattery, W. L. & Cameron, A. G. W. in *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. J.) 617–620 (Lunar Planet. Inst., Houston, 1986).
8. Kipp, M. E. & Melosh, H. J. in *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. J.) 643–647 (Lunar Planet. Inst., Houston, 1986).
9. Melosh, H. J. & Kipp, M. E. *Lunar Planet. Sci. Conf.* **XX**, 685–686 (1989).
10. Thompson, S. L. & Lauson, H. S. Rep. SS-RR-71 0714 (Sandia Natl Lab., Albuquerque, 1972).
11. Tillotson, J. H. Rep. GA-3216 (General Atomic, San Diego, 1962).

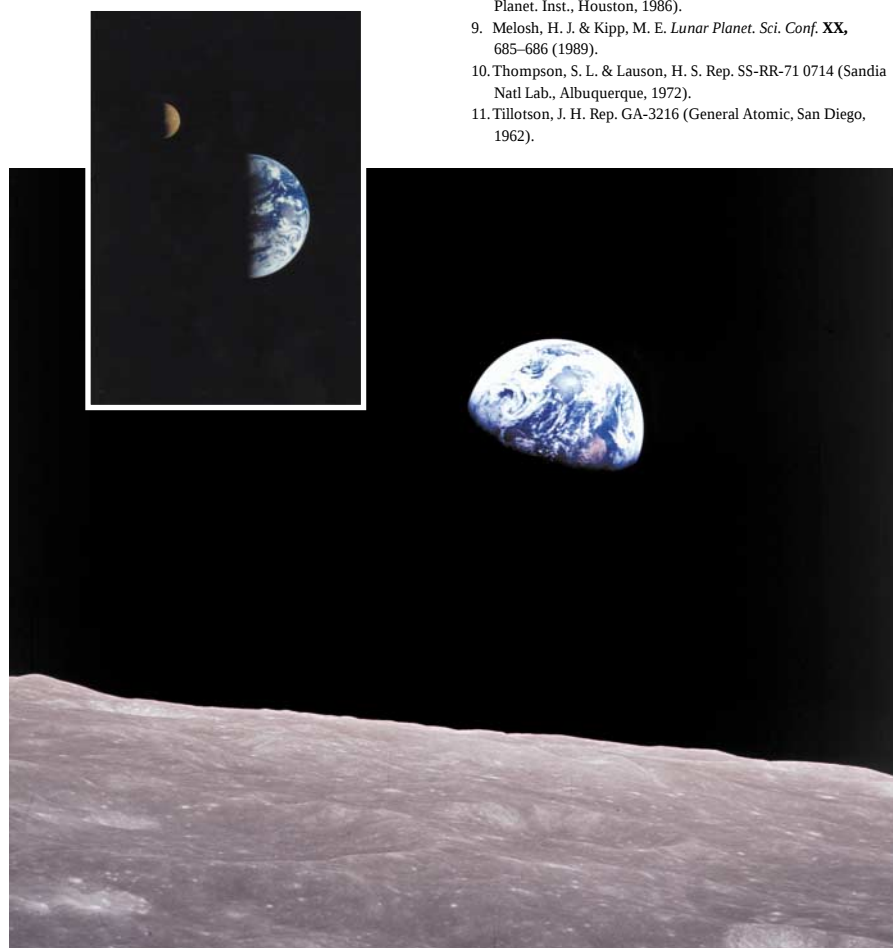


Figure 1 Earth and Moon — shock simulation.

Obituary

Donald J. Cram (1919–2001)

Donald Cram, a gifted and creative organic chemist, died peacefully at his home in Palm Desert, California, on 17 June. He had been plagued periodically for 30 years by metastatic melanoma, a disease to which he succumbed rapidly in the end.

Cram was born in semi-rural Brattleboro, Vermont, on 22 April 1919. His father, a successful attorney, died when Cram was four years of age. He and his four sisters were raised by their mother, a woman of considerable resolve. When he was 16 years of age, the family dispersed and Cram became self-supporting, working first in Florida and then on Long Island, New York. This early independence, acquired while both working and attending high school, instilled the personal characteristics and standards that guided him so well throughout his life.

In 1937, Cram obtained a full four-year scholarship to Rollins College near Orlando, Florida, which enabled him to pursue chemistry as a major. There he began a lifelong love affair with organic chemistry, a subject that not only intrigued him intellectually but also offered an escape from the tedium of the low-paid, repetitive tasks that he had come to loathe. Cram's strongest desire was to become a university professor because no other profession offered the freedom to choose research problems and gamble that the choices were correct. He started by taking a master's degree in 1942, under Norman H. Cromwell at the University of Nebraska. The United States had just entered the Second World War, and Cram was then hired by the Merck pharmaceutical company to work on the development of penicillin, which he did until the war's end.

In 1945, Cram moved to Harvard University where, in a mere 18 months, he completed his PhD under Louis Fieser. He was then awarded an American Chemical Society fellowship, which allowed him to carry out independent postdoctoral research at an institution of his choice. Cram knew the eastern United States well, so he went for the unknown — the University of California at Los Angeles. The proximity of the Pacific Ocean on the one hand, and ski slopes on the other, no doubt played a part in this choice. UCLA was then only just beginning to develop as a research university, but a world-class organic chemistry group had already been established there by Saul Winstein, William G. Young and Thomas Jacobs. In

1948, Cram accepted UCLA's offer of an assistant professorship and he immediately began to realize his dream — that of setting up his own research group. UCLA was to be his scientific home for the next 50 years.

The last half of the twentieth century may prove to have been the golden age of organic chemistry, and Cram's research career flourished. Major topics in the earliest part of this period were the elucidation of organic reaction mechanisms, the development of new synthetic methods, and the construction of molecules designed to test ever-advancing concepts of structure and bonding. Cram's career touched on all of these subjects, and two themes pervade his chemistry: the application of stereochemistry and chirality as research tools, and the design of molecular architecture for particular functions. He was never without a satchel filled with molecular models, attesting to his single-minded dedication to this last theme.

There are six areas of chemistry to which Cram made seminal contributions or which he created outright. Each represents a change of interest, occurring at roughly ten-year intervals. These moves stemmed from Cram's view that more than a decade of intense work on any subject would exhaust his creativity in that area.

Those areas and his achievements, in chronological order, were the isolation and structural characterization of natural products; the discovery of phenonium ions and stereochemical study of the mechanism of the Wagner–Meerwein rearrangement, triggered by his investigation of the acid-catalysed racemization of a degradation product of the antibiotic citrinin; the design and synthesis of the cyclophane hydrocarbons and their derivatives, in which the π -electron systems of aromatic rings were forced into face-to-face contact, causing perturbations of their chemical and physical properties; the stereochemical and mechanistic study of electrophilic substitution reactions at saturated carbon centres, which later became identified with carbanion chemistry; the discovery and development of host–guest

complexation chemistry and its relevance to biology, for which Cram was awarded the 1987 Nobel Prize in Chemistry along with Jean-Marie Lehn and Charles Pedersen; and, finally, the design and synthesis of carcerands, or 'molecular bottle' species, which encapsulate reactive molecules that can subsequently be converted within the carcerand to notoriously unstable species such as cyclobutadiene and *o*-benzyne.

Entrapment in the molecular vessel allowed these and similar species to be studied as never before. Cram himself considered bottle molecules to be his finest scientific discovery.

Cram's passion for chemistry was matched by his participation in sports requiring skill, physical exertion, and often courage. He could be found on the ski slopes of southern California during the winter, and surfing with his friends — a group known as 'the old guys' — at San Onofre beach in the summer. He had a lifelong penchant for singing ballads, accompanying himself on the guitar. He continued to take part in sporting activities until his eightieth year, and had recently taken up golf.

Above all, Donald Cram was a strong and realistic man who observed his own approaching death with calm and with gratitude for having had all of his dreams fulfilled. He is survived by his third wife, Caroline.

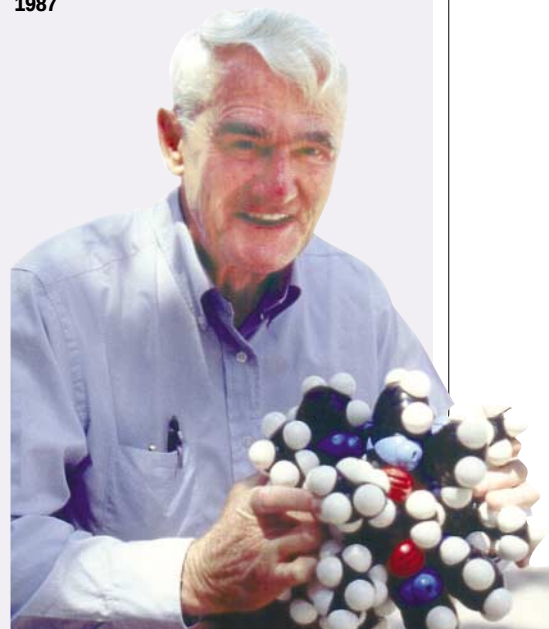
M. Frederick Hawthorne

M. Frederick Hawthorne is in the Department of Chemistry and Biochemistry, University of California at Los Angeles, California 90095-1569, USA.

e-mail: mfh@chem.ucla.edu

Web link

♦ <http://www.nobel.se/chemistry/laureates/1987>



Pioneer in organic chemistry, especially host–guest complexes

Finer features for functional microdevices

Micromachines can be created with higher resolution using two-photon absorption.

Compared with light or electron-beam lithography, the virtue of two-photon photopolymerization¹ as a tool for making microdevices lies in its three-dimensional capability, which has found application in photonic devices^{2–4} and micromachines^{1,5,6} with feature sizes close to the diffraction limit. Here we show that the diffraction limit can be exceeded by nonlinear effects to give a subdiffraction-limit spatial resolution of 120 nanometres. This allows functional micromachines to be created and shifts the working wavelength of photonic and opto-electronic devices into the visible and near-infrared region.

Commercially available resin (SCR500; JSR, Japan), consisting of urethane acrylate monomers and oligomers as well as photoinitiators, is transparent to an infrared laser and allows it to penetrate deeply. The resin can be photopolymerized by using two-photon absorption (TPA)^{7–9} to create three-dimensional structures. By pinpoint-scanning the laser focus according to pre-programmed patterns, designs can be faithfully replicated to matter structures.

Figure 1a–f shows scanning electron micrographs of ‘micro-bull’ sculptures. These 10- μm -long, 7- μm -high bulls are the smallest model animals ever made artificially, and are about the size of a red blood cell. The tiny volume attainable for such micromachines would allow them to be transported to locations inside the human body through even the smallest blood vessels, for example to deliver clinical treatments.

To depict the tiny features of our microbulls, a fabrication accuracy of about 150 nm was needed, which we achieved by using nonlinear processes of the photochemical reactions involved. Polymeriza-

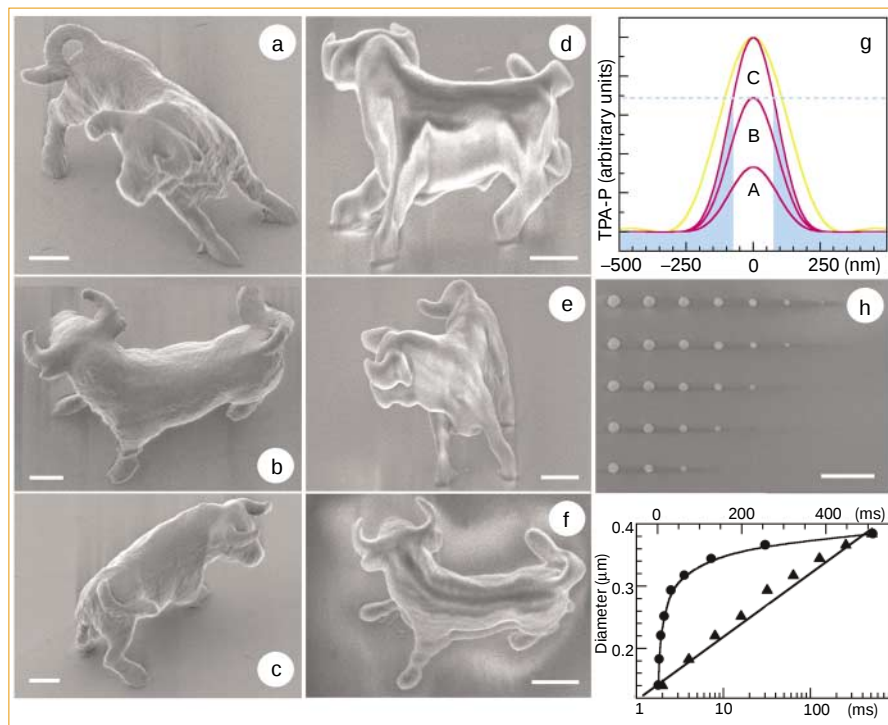


Figure 1 Microfabrication and nanofabrication at subdiffraction-limit resolution. A titanium sapphire laser operating in mode-lock at 76 MHz and 780 nm with a 150-femtosecond pulse width was used as an exposure source. The laser was focused by an objective lens of high numerical aperture (~ 1.4). **a–c**, Bull sculpture produced by raster scanning; the process took 180 min. **d–f**, The surface of the bull was defined by two-photon absorption (TPA; that is, surface-profile scanning) and was then solidified internally by illumination under a mercury lamp, reducing the TPA-scanning time to 13 min. **g**, Achievement of subdiffraction-limit resolution, where A, B and C respectively denote the laser-pulse energy below, at and above the TPA-polymerization threshold (dashed line). The yellow line represents the range of single-photon absorption. TPA-P, TPA probability. **h**, Scanning electron micrograph of voxels formed at different exposure times and laser-pulse energies. **i**, Dependence of lateral spatial resolution on exposure time. The laser-pulse energy was 137 pJ. The same data are presented using both logarithmic (triangles; bottom axis) and linear (circles; top axis) coordinates, to show the logarithmic dependence and threshold behaviour of TPA photopolymerization. Scale bars, 2 μm .

tion occurred only in the vicinity of the focal spot and the size of solidified voxels (three-dimensional volume elements) was reduced because of the quadratic depen-

dence of TPA probability on the photon fluence density. Furthermore, photogenerated radicals in this volume were subject to scavenging by dissolved oxygen molecules¹⁰, so polymerization reactions were not initiated and propagated if the exposure energy was less than a critical value.

This property defined a TPA threshold and excluded the low-intensity lobe (the zero-order edge and the subsidiary maxima of the Airy pattern) from polymerization and thus further reduced the voxel size (Fig. 1g). The diffraction limit imposed by Rayleigh criteria was simply a measure of focal-spot size, and did not put any restraint on the actual size of solidified voxels. We were thus able to design voxels of any small size by choosing an appropriate laser-pulse energy and exposure time, because only the region with energy above the TPA threshold was solidified (Fig. 1h).

A logarithmic dependence of voxel size on exposure time (Fig. 1i) results from the exponential decay of the oligomer/monomer

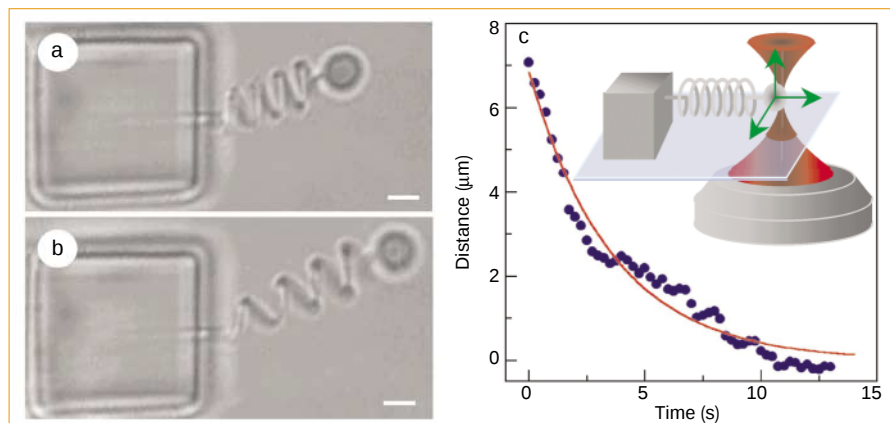


Figure 2 Functional micro-oscillator system, in which not only the spring but also the cubic anchor and the bead were produced using our two-photon absorption system. The oscillator was kept in ethanol so that the buoyancy would balance gravity and eliminate bead–substrate friction. **a**, **b**, The spring in its original (**a**) and extended (**b**) states. Scale bars, 2 μm . **c**, Restoring curve of the damping oscillation; inset, diagram showing driving of the oscillator by using laser trapping.

concentration during exposure. Our spatial resolution of 120 nm is superior to that achieved by conventional rapid-laser prototyping and by conventional TPA fabrication^{1–8} (smaller voxels can be formed, but it is difficult for isolated voxels to appear in the same scanning electron micrograph; in an actual fabrication, the spatial resolution may be better than 120 nm).

As an example of subdiffraction-limit fabrication, we produced a micro-oscillator, which must be the smallest functional micromechanical system produced (note that the microspring shown in Fig. 2a, b has a spiral diameter of only 300 nm). To operate such a minute spring, we converted it into an oscillator by fixing one end to an anchor attached to a glass substrate and polymerizing a bead (diameter, 3 µm) at the other end. We used laser-trapping force^{11,12} to capture the bead, pulled the spring (Fig. 2b), and then released it from its displacement (Fig. 2c, inset) to set the vibration in motion (see movie in supplementary information).

An 820-nm, 19-mW laser offered a trapping force of about 3 piconewtons, which is equivalent to a 20g acceleration of the bead, where g is the acceleration due to gravity. However, because of the large specific surface of the oscillator, viscosity heavily damped the oscillation. We assume that the damping force is proportional to the velocity of the bead movement, $f_{vis} = 6\pi\eta rv$ (Stokes' law), where η (1.084×10^{-3} pascals

per second at 25 °C) is the liquid viscosity and v is the bead velocity. The spring damping oscillation can be expressed as $m d^2x/dt^2 = -kx - 6\pi\eta r(dx/dt)$, where m (1.6×10^{-14} kg) is the bead mass and k is the spring constant to be determined. Figure 2c shows the bead displacement during spring restoration against time, from which the spring constant was deduced to be 8.2 nN m^{-1} . Such soft springs are directly applicable to the investigation of the mechanical properties of micro-objects.

Satoshi Kawata, Hong-Bo Sun,

Tomokazu Tanaka, Kenji Takada

Department of Applied Physics, Osaka University,
Suita, Osaka 565-0871, Japan

e-mail: kawata@ap.eng.osaka-u.ac.jp

1. Maruo, S., Nakamura, O. & Kawata, S. *Opt. Lett.* **22**, 132–134 (1997).
2. Sun, H.-B., Matsuo, S. & Misawa, H. *Appl. Phys. Lett.* **74**, 786–788 (1999).
3. Cumpston, B. H. *et al. Nature* **398**, 51–54 (1999).
4. Sun, H.-B. *et al. Appl. Phys. Lett.* **79**, 1–3 (2001).
5. Maruo, S. & Kawata, S. *IEEE J. Microelectromech. Syst.* **7**, 411–415 (1998).
6. Sun, H.-B. *et al. Opt. Lett.* **25**, 1110–1112 (2000).
7. Parthenopoulos, D. A. & Rentzepis, P. M. *Science* **245**, 843–845 (1989).
8. Denk, W., Strickler, J. H. & Webb, W. W. *Science* **248**, 73–76 (1990).
9. Wu, E. S., Strickler, J. H., Harrell, W. R. & Webb W. W. *Proc. SPIE* **1674**, 776–782 (1992).
10. Flory, P. J. *Principals of Polymer Chemistry* (Cornell Univ. Press, New York, 1952).
11. Ashkin, A. *Phys. Rev. Lett.* **24**, 156–159 (1970).
12. Inouye, Y., Shoji, S., Furukawa, H., Nakamura, O. & Kawata, S. *Jpn J. Appl. Phys.* **37**, L684–L686 (1998).

Supplementary information is available on Nature's website (<http://www.nature.com>).

Sex determination

Viviparous lizard selects sex of embryos

No one suspected that temperature-dependent sex determination (TSD)^{1–3}, whereby the sex of embryos depends on the temperature at which they develop, might occur in viviparous (live-bearing) reptiles, because thermoregulation in the mother results in relatively stable, raised gestation temperatures. But here we show that developing embryos of the actively thermoregulating viviparous skink *Eulamprus tympanum* are subject to TSD, offering the mother the chance to select the sex of her offspring and a mechanism to help to balance sex ratios in wild populations.

Sex determination is the programmed cascade of events through which an undifferentiated gonad develops into a testis or an ovary. In vertebrates, sex is determined either by a genotypic mechanism at the time of fertilization, which depends only on genetic factors, or by environmental factors that act after fertilization. Species that are subject to TSD provide an example of the latter mechanism and usually lack heteromorphic sex chromosomes. Reptiles rely on

either temperature or genetic factors to influence the sex of their offspring⁴.

E. tympanum is a medium-sized scincid lizard found in high-elevation habitats in southeastern Australia, with a litter size of one to five young⁵. As no species within *Eulamprus* has detectable heteromorphic sex chromosomes, we investigated whether *E. tympanum* might be subject to TSD. We maintained mothers at different laboratory temperatures and used palpation of the hemipenes⁶ and histology of neonatal gonads^{7,8} to establish sex. To our surprise, we discovered that gestation temperature has a highly significant effect on sex ($P < 0.001$), with warmer temperatures giving rise to male offspring (Fig. 1).

Active thermoregulation by pregnant viviparous lizards distinguishes the thermal environment of development from that in oviparous species. A combination of active thermoregulation and TSD provides the female lizard with the opportunity to select the sex of her offspring. In the laboratory, all females provided with unlimited conditions for thermoregulation maintained body temperatures of 32 °C and produced exclusively male offspring. Equal sex ratios resulted from natural gestation in two field seasons.

We do not understand the mechanism by

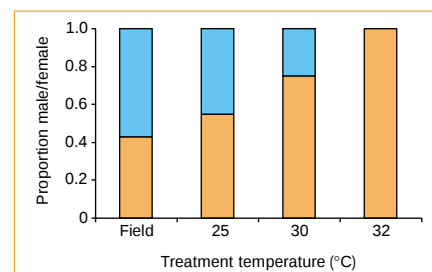


Figure 1 Influence of gestation temperature on the sex ratio of offspring of the viviparous lizard *Eulamprus tympanum*. Females maintained at 32 °C ($n = 21$) for the duration of pregnancy gave birth to exclusively male offspring ($n = 55$); those maintained at 30 °C ($n = 20$) gave birth to predominantly male offspring ($n = 58$; 75% were male); those maintained at 25 °C ($n = 11$) gave birth to offspring of both sexes ($n = 20$; 55% were male); those undergoing most of their gestation in the field ($n = 24$) also produced a mix of sexes ($n = 58$; 43% were male). Orange portions of bars represent male offspring; blue portions represent female offspring.

which females select body temperature to give equal sex ratios in the field; the implication is that thermoregulatory conditions may be restricted. Alternatively, other factors, such as unbalanced adult sex ratios, may result in mothers selectively thermoregulating to produce offspring that help to balance the population sex ratio. A viviparous skink from Tasmania adjusts the sex ratio of its offspring according to the operational sex ratio of the adult population⁹; presumably, TSD provides the mechanism for selection of neonatal gender by the mothers. Our population of *E. tympanum* has an adult sex ratio that is not significantly different from unity¹⁰ in the field, where they produce an equal sex ratio of neonates; however, our laboratory population was all female and produced all sons when given the opportunity to thermoregulate.

TSD may explain the fact that *E. tympanum*, like many other viviparous taxa, is restricted to alpine regions. The warmer temperatures further down the slopes would encourage production of exclusively male offspring and lead to the eventual extinction of those populations. A combination of alpine distribution and TSD is likely to be a problem in the event of rapid climate change or global warming, as these species may not be able to evolve rapidly enough to compensate¹¹. For alpine species, there can be no retreat to cooler climates, so a rise in environmental temperature would result in increased production of males. Models predict a temperature rise of 4 °C by 2100 (ref. 12), which could seriously alter the sex ratio and lead to extinction of species such as *E. tympanum*.

Kylie A. Robert, Michael B. Thompson

School of Biological Sciences and Institute of
Wildlife Research, University of Sydney,
New South Wales 2006, Australia
e-mail: kr robert@bio.usyd.edu.au

1. Bull, J. J. Q. *Rev. Biol.* **55**, 3–21 (1980).

2. Janzen, F. J. & Paukstis, G. L. *Q. Rev. Biol.* **66**, 149–179 (1991).
3. Cree, A., Thompson, M. B. & Daugherty, C. H. *Nature* **375**, 543 (1995).
4. Ciofi, C. & Swingland, I. R. *App. Anim. Behav. Sci.* **51**, 251–265 (1997).
5. Cogger, H. G. *Reptiles and Amphibians of Australia* 3rd edn (Reed, Sydney, 1986).
6. Harlow, P. S. *Herpet. Rev.* **27**, 71–72 (1996).
7. Yntema, C. L. *J. Morphol.* **167**, 297–304 (1981).
8. Yntema, C. L. & Mrosovsky, N. *Herpetologica* **36**, 33–36 (1980).
9. Olsson, M. & Shine, R. *J. Evol. Biol.* **14**, 120–128 (2001).
10. Schwarzkopf, L. & Shine, R. *Oecologia* **88**, 562–569 (1991).
11. Janzen, F. J. *Proc. Natl Acad. Sci. USA* **91**, 7487–7490 (1994).
12. Whetton, P. H. *et al. Climate Res.* **16**, 181–201 (2001).

Ancient chronology

Astronomical orientation of the pyramids

Spence speculates that Egypt's pyramid builders found true north by using a plumb line: when the stars Kochab and Mizar were seen on the same vertical, one was facing north¹. As evidence in support of this hypothesis, she points to the proposed interstar-line precession past the north celestial pole at a rate of 27' per century (cy). We argue that a mathematical error affects this result, which when corrected points more strongly to a different pair of stars. This suggests that the conventional ancient chronology, instead of being compressed, may actually have to be expanded slightly.

The elementary error here is that the interstar-line drift of 27' cy⁻¹ occurs at an altitude of 30° (Giza's latitude), but to apply this drift to ground orientation, one must divide by cosine 30°. So, on the ground, 31' cy⁻¹ is the actual misorientation rate. Thus, the actual drift rate is significantly greater than Spence's empirical rate of orientation change of the pyramids themselves (the slope of line *a* in Fig. 4 of ref. 1).

When the pyramids were built, only two stars brighter than fifth magnitude lay anywhere near the pole: Thuban (3.65 mag, 1.5° distant) and the irregularly variable star 10 Draconis (4.5–5.0 mag). In 2627 BC, the pole was equidistant (1°) from each star, so the pole was the obtuse apex of a squat isosceles triangle formed between itself and the two stars. When both stars were at the same altitude, north was the direction bisecting them. (For more than a millennium after 2627 BC, there was no star brighter than 10 Draconis nearer to the celestial pole.)

Among several mechanical methods, north could have been determined in the dark by sighting the two horizontal stars simultaneously against a pointed post, the pyramidal top of an obelisk, or any similar object; when the observer can eclipse both stars simultaneously on opposite sides of the peak, the line from the observer to the peak points northwards. This simultaneous-eclipse method does not require the post or obelisk to be illuminated, making it simpler than Spence's plumb-line method;

there is no easy way to see a plumb line at night while retaining the observer's night-vision acuity.

Although 10 Draconis is barely visible under modern industrial skies, it is recorded in all four large preclassical naked-eye star catalogues: Hipparchos, 128 BC; Ulugh Beg, AD 1437; Tycho, AD 1601; and Hevelius, AD 1661. Spence's Kochab and Mizar are indeed brighter than Thuban, but the eye-precision she assumes implies that the Kochab–Mizar line will confusingly pass into detectable and uncentred non-verticality in a matter of a few (perhaps ten) seconds. (Spence's suggestions of 5-year or 1–2-year precision for dating the pyramids imply a surveying precision of about 1'.) So Spence's method, although possible, would require agile quickness. In contrast, the midpoint between Thuban and 10 Draconis gives a ground orientation within 1' of true north for over 5 minutes on either side of its transit. The very slow motion of these stars (and the small size of any potential orientation error from their use) is due to their close proximity to the celestial pole.

The precision of raw, naked-eye stellar observations can be significantly better than 3', but we justify the utility of our two stars by reference to the scrupulous naked-eye catalogues of Tycho and Hevelius^{2,3}. Tycho's raw data survive for both stars⁴, eight observations in all: r.m.s. errors are 2' for Thuban and 3' for 10 Draconis. In Hevelius' catalogue, the equatorial coordinates of Thuban and 10 Draconis (his Draco stars 8 and 32) have great-circle errors of 1' and 0', respectively. Thus, the dimness of 10 Draconis was in itself no barrier to accurate measurement of its position in pre-industrial times, and such precision could easily be replicated for an azimuth observation, even using simple instruments, by positioning the observer's eye at a large enough distance from the eclipsing post.

In 2627 BC, the misorientation associated with our obvious and straightforward method was null but precessionally increasing at 27.4' cy⁻¹ in azimuth, which matches Spence's 28' cy⁻¹ empirical rate much more closely than her Mizar–Kochab method (31' cy⁻¹). This implies dates of 2638 BC for Khufu's pyramid and 2607 BC for Khafre's. (Error estimates could be 2–10 years, depending on assumptions regarding the builders' craftsmanship.) These dates are a few decades outside the conventional ranges Spence cites⁵. But our implied date for the ascension of Khufu (2640 BC) is twice as near to a conventional boundary as Spence's 2480 BC (Table 1 in ref. 1). Back-disparity makes more sense than Spence's very forward dates, when current orthodoxy is based on king lists that are "seldom complete"¹.

It seems odd that either method would have been used before the time when it was correct. Because the best pyramid orienta-

tions occur for the two greatest pyramids, this could simply indicate that engineering science peaked at the time of Khufu–Khafre. Thuban passed within 0.1° of the pole in 2800 BC, a chance event that may have stimulated the historical flowering of celestially based surveying, which was unquestionably used for the pyramids built soon after that at Giza. A stellar explanation of the Giza pyramids' location (in latitude) has already been proposed⁶.

Latitude (but not orientation) could be found in a single night near the time of the winter solstice anywhere close to 2600 BC by bisecting Thuban's circumpolar semicircle, because this star was within 10° of the equinoctial colure for centuries after 2700 BC. Because 10 Draconis was within 10° of the solstitial colure for three decades either side of 2613 BC, on the same or any neighbouring night, orientation might also have been found by bisecting the circumpolar arc of 10 Draconis. Although twilight would have cut the arc to slightly less than 180°, this still would have been adequate for the purpose. By coincidence, both orientation methods that depended on 10 Draconis were most accurate at virtually the same time: 2627 BC for the Spence interstar method applied to Thuban and 10 Draconis, and 2613 BC for the 10 Draconis arc method.

Before Spence's proposal¹, a possible connection was suggested⁷ between precession and the pyramids' misorientation, but the horizon-based observations proposed would be too prone to difficulties with high refraction, uneven topography, dip and atmospheric extinction to be practical. So, despite a few disagreements, we welcome Spence's creativity in pointing out the possibilities of orienting the pyramids by observing northern stars higher in the sky and near to the meridian, which doubly minimizes corruption by refraction.

Dennis Rawlins*, Keith Pickering†

**DIO, The International Journal of Scientific History*, Box 19935, Baltimore, Maryland 21211-0935, USA

†*Analysts International Corporation*, 3601 West 76th Street, Minneapolis, Minnesota 55435, USA
e-mail: keithp@minn.net

1. Spence, K. *Nature* **408**, 320–324 (2000).
2. Rawlins, D. *Isis* **73**, 259–265 (1981).
3. Rawlins, D. *DIO Int. J. Sci. Hist.* **3**, 3–106 (1993).
4. Dreyer, J. L. E. (ed.) *Tychonis Brahe Dani Opera Omnia* Vol. 11, 382 and Vol. 12, 170 (Societatis Linguae et Litteraturae Danicarum, Copenhagen, 1913–29).
5. von Beckerath, J. in *Chronologie des Pharonischen Ägypten* 188 (von Zabern, Mainz, 1997).
6. Rawlins, D. *Vistas Astron.* **28**, 255–268 (1985).
7. Haack, S. C. *J. Hist. Astron.* **7** (suppl.), 119–125 (1984).

Spence replies — Rawlins and Pickering have correctly identified an error: I should indeed have divided the calculated figures for the distance of the line between β Ursae Minoris and ζ Ursae Majoris from the pole

by cosine 30° to convert the rate of drift (which is correct) into azimuth. However, this error does not invalidate my method¹: the increase in the slope is small and the revised gradient fits well with the data originally presented (Fig. 1). The decision to discard the alternative pair of stars first considered is also now more convincing, as their revised drift is around $37' \text{ cy}^{-1}$, which cannot be accommodated by the archaeological data. The revision therefore does not affect my results¹, apart from being more compatible with a shorter reign for Snofru.

Rawlins and Pickering also question whether orientations could be achieved during the short period when the two stars are in simultaneous transit. This would not be a problem as, assuming that the plumb line was hung from a frame, the sighting device could be adjusted to keep the plumb line equidistant between the two stars for several minutes as they came into simultaneous transit. The alignment could also have been checked on successive nights to ensure that it was exact. The fact that this method has a short and clearly defined period (the stars move rapidly out of plumb in opposite directions), within which the alignment is established, is one of the reasons why it is such an accurate method of orientation. The stars used are very bright and low light on a plumb line would not have substantially affected visibility.

I find the alternative method proposed by Rawlins and Pickering unconvincing. 10 Draconis is very faint, barely visible to the naked eye on a clear, dark night. The fact that a star appears in historical star catalogues from 128 BC onwards does not mean that it was used or even recognized by the ancient Egyptians around 2500 BC. These historical star catalogues were compiled within a context of interest in detailed astronomical measurement and recording, for which we have no evidence from ancient Egypt. Establishing when two stars that are so close and faint are exactly horizontal and simultaneously eclipsed would not be easy. In addition, if the Egyptians thought that the pole was equidistant from two stars, both of which are close to the pole and one barely visible, it is unclear why they did not simply use the brighter star (α Draconis) in a bisection method.

Haack has also suggested a link between the misorientation of pyramids and precession², proposing that the pyramids were oriented towards the rising or setting of a star on the east or west horizon. Variable horizon conditions between sites, and even on the same site (the ground slopes upwards to the west at Giza, precluding observations equivalent to those made to the east), coupled with all the problems involved with observing stars that are close

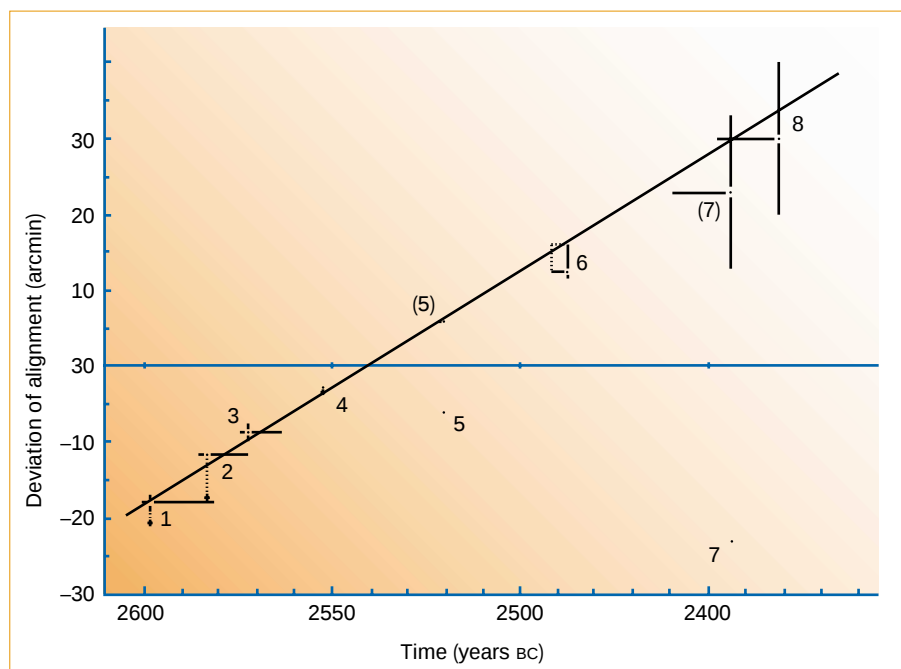


Figure 1 Astronomical modelling of the simultaneous-transit method of orientation. The points are plotted from the data given in Table 1 of ref. 1. The corrected gradient for measurements of orientation using the simultaneous transit of β Ursae Minoris and ζ Ursae Majoris is plotted over the archaeological data. The point at which this line crosses zero can be recalibrated to 2467 BC by astronomical modelling. 1, Meidum; 2, Bent Pyramid; 3, Red Pyramid; 4, Khufu; 5, Khafre; 6, Menkaure; 7, Sahure; 8, Neferirkare. Numbers in parentheses denote points replotted with positive rather than negative values to conform to the dominant trend of the alignments.

to the horizon, make this method untenable. However, as both alternative methods can be modelled to produce results that are comparable with the archaeological data, the context must also be considered. The method I propose and the stars used are consistent with other evidence of Egyptian astronomy at this time.

In Khufu's pyramid, four narrow shafts run north and south at an upward angle from two of the inner chambers. These shafts are probably oriented towards the culminations of important stars^{3–5}, although this is not accepted by all Egyptologists. It has been suggested that one of the northern shafts is oriented towards the culmination of β Ursae Minoris⁵, which is also one of the stars used in the simultaneous-transit method. Other investigations (T. van Albada and A. Egberts, personal communication) confirm the orientation towards β Ursae Minoris and give dates that are compatible with my results. Independent archaeological data for the shafts and the bases suggest that both are aligned towards culminations, that both involve β Ursae Minoris and that both require a forward shift of the chronology of the period.

Later pictorial evidence, such as the Senenmut astronomical ceiling, suggests that the Egyptians envisaged a line running between two circumpolar constellations as a method of finding north⁶. One of these constellations is definitely Ursa Major⁶, the other is probably Ursa Minor⁷. Later textual evidence suggests the use of a plumb line

and Ursa Major⁶.

The Egyptians were almost certainly unaware that their orientation method was precession-dependent and that it was therefore initially inexact, otherwise they would have used a more accurate bisection method. There is no evidence from any period of their history that the ancient Egyptians measured or understood the concept of latitude, which emerged during Hellenistic times within the context of a far more sophisticated programme of astronomical observation and investigation. Rawlins' speculation⁸ of a stellar explanation of the Giza pyramids' location in terms of latitude is unlikely, as it accounts for the location of only three of the many kings' pyramids in Egypt, and the pyramid of Djedefre, which was constructed between the reigns of Khufu and Khafre, was built at a different site.

Kate Spence

Faculty of Oriental Studies, University of Cambridge, Cambridge CB3 9DA, UK
e-mail: kes1004@cam.ac.uk

1. Spence, K. *Nature* **408**, 320–324 (2000).
2. Haack, S. C. J. *J. Hist. Astron.* **15** (suppl.), S119–S125 (1984).
3. Badawy, A. *Mitteilungen des Instituts für Orientforschung der Deutschen Akademie der Wissenschaften zu Berlin* **10**, 189–206 (1964).
4. Trimble, V. *Mitteilungen des Instituts für Orientforschung der Deutschen Akademie der Wissenschaften zu Berlin* **10**, 183–187 (1964).
5. Bauval, R. G. & Gilbert, A. G. *Disc. Egyptol.* **28**, 5–13 (1994).
6. Zába, Z. *L'orientation astronomique dans l'ancienne Egypte et la précession de l'axe du monde* (Editions de l'Académie Tchecoslovaque des Sciences, Prague, 1953).
7. Davies, V. L. J. *J. Hist. Astron.* **16** (suppl.), S102–S104 (1985).
8. Rawlins, D. *Vistas Astron.* **28**, 255–268 (1985).

Structural mimicry in bacterial virulence

C. Erec Stebbins & Jorge E. Galán

Section of Microbial Pathogenesis, Boyer Center for Molecular Medicine, Yale School of Medicine, New Haven, Connecticut 06536, USA

An important mechanism underlying the strategies used by microbial pathogens to manipulate cellular functions is that of functional mimicry of host activities. In some cases, mimicry is achieved through virulence factors that are direct homologues of host proteins. In others, convergent evolution has produced new effectors that, although having no obvious amino-acid sequence similarity to host factors, are revealed by structural studies to display mimicry at the molecular level.

The pressures of survival have engendered a fascinating spectrum of adaptations in organisms. Different organisms have evolved sophisticated methods to exploit the surrounding environment and each other. An important mechanism that frequently reoccurs in this process of adaptation is that of mimicry. Many organisms, both large and small, have found a selective advantage in imitating the appearance or function associated with an otherwise distinct creature or aspect of the natural environment. In many of these cases organisms imitate or copy the appearance of something else, either for the purposes of concealment (as with the African praying mantis and chameleons) or to send a message (for example, coloration in non-poisonous species to mimic poisonous ones). Some of the most interesting examples of mimicry, however, may occur in the microscopic world. Recent studies have begun to reveal that many bacterial pathogens mimic the function of host proteins to manipulate host physiology and cellular functions for the microbe's benefit^{1–8}. This is in contrast with the strategies used by some pathogens that involve microbial products with activities lacking clear counterparts in eukaryotic cells^{9–11}. Here we consider recent structural work that provides unique insights into the mechanisms of host mimicry by bacterial virulence factors. In some cases, these factors are homologues of host proteins that have been incorporated into the genome of the pathogen and subverted for its benefit. In others, convergent evolution has produced new effectors that have no obvious relationship to host factors. However, although hidden at the sequence level, the determination of the crystal structures of several bacterial factors and bacterial–host protein complexes has revealed the presence of mimicry at the molecular level. Examination of such factors is providing important insights into the interplay between host and pathogen, the mechanisms underlying eukaryotic functional homologues, and the nature of the evolutionary dynamics shaping these complex ecologies.

Evolutionary mechanisms for mimicking the host

The ability to modulate cellular activities of the host at the molecular level through functional mimicry is a powerful tool for a bacterial pathogen. It allows the bacterium to be precise and limited in its effects, which can be useful in achieving its goals (for example, internalization into a host cell through limited disruption). However, obtaining virulence factors with such activity presents a daunting challenge. A pathogen may acquire such effector molecules by either obtaining 'foreign' genes through horizontal transfer (in particular, host protein homologues), or through the process of convergent evolution. Mimicry through convergent evolution is perhaps the more intriguing of the two, as it involves taking 'materials' (genes and the proteins that they encode) already available to the pathogen and then 'sculpting' them to perform a new function. Such a protein would usually have a distinct three-dimensional architecture from that of the molecule it mimics, but would typically have evolved to imitate the chemical groups on the surface of its functional homologue.

For many of the functional mimics used by bacterial pathogens, the mimicry is indeed achieved through homologous enzymes that have been subverted for the benefit of the pathogen^{1,6,7,12}. These enzymes are often easily identifiable by sequence alignments as they contain highly conserved active sites or regulatory motifs (Table 1). These bacterial homologues of host proteins often differ from their host counterparts through alterations in substrate specificity, absence of regulatory control domains, and/or modulation of their intrinsic activity.

Examples of virulence factors with such properties can be found among the array of effector proteins that many pathogenic bacteria inject into host cells through a specialized organelle termed the type III secretion system¹³. Two such proteins are the tyrosine phosphatases YopH and SptP from *Yersinia* spp. and *Salmonella* spp., respectively^{1,5,14,15}. As tyrosine phosphorylation does not commonly occur in bacteria, it is probable that these molecules have specifically evolved to modulate host cellular functions. This is supported by their effects on eukaryotic cells. YopH, for example, disrupts focal adhesions by dephosphorylating p130^{cas} and the focal adhesion kinase (FAK), leading to a paralysis of macrophage attack on the bacterium^{16,17}. Although the substrates of the tyrosine phosphatase activity of SptP have not been identified, this protein is involved in the reversion of the cellular responses stimulated by *Salmonella* to preserve the fitness of its intracellular niche⁴ (see below). Both YopH and the carboxy-terminal half of SptP possess sequence similarity to eukaryotic tyrosine phosphatases, and the crystal structures of these molecules show that they share a very similar fold with these enzymes, particularly in the active site^{5,14}. Other examples are the serine threonine kinase YpkA (ref. 6) and the cysteine protease YopJ from *Yersinia* spp.¹⁸, and also the inositol polyphosphatase SopB from *Salmonella* spp.⁷ (Table 1). Although the structures of these molecules are unknown, similarities with eukaryotic counterparts are detectable at the level of primary amino-acid sequence. Therefore, in these cases, the pathogens have recruited virulence determinants with sequence and structural similarities to host enzymes, but with unique substrates and regulation adapted to the needs of the pathogen.

Host mimicry through convergent evolution

Many bacterial virulence factors with activities similar to host enzymes do not show any sequence similarity to eukaryotic proteins^{2,419–21}. As structural and functional similarities can occur even in the absence of sequence similarity, it was unclear how this type of virulence factor functioned. However, the solution of several crystal structures of such factors and their complexes with host targets has helped to provide valuable insights on this process^{3,5,22}. In some cases these bacterial mimics possess a structural architecture (the fold) that differs markedly from that of their host functional homologues. However, the molecular surfaces that interact with their targets, the true level at which natural selection ultimately sculpts, are seen as excellent mimics of proteins that operate normally in the cell. We consider two specific examples where

structural information has been a central element in understanding the function and evolution of a virulence factor.

Salmonella SptP and mimicry of signal transduction effectors

The enteric pathogen *Salmonella* delivers into the host cell two highly related bacterial proteins (SopE and SopE2) through a specialized organelle termed the type III secretion system²³. These proteins function as guanine nucleotide exchange factors (GEFs) that activate Rac1 and Cdc42 (refs 2 and 24). Activation of these GTPases leads to profuse rearrangements of the actin cytoskeleton and subsequent bacterial internalization into intestinal epithelial cells. Once safely inside the cell, *Salmonella* actively contributes to the restoration of the normal architecture of the host cell cytoskeleton by delivering another effector protein, SptP, thereby preventing any potential harm to its protected niche resulting from excessive Rho GTPase signalling⁴. The amino-terminal half of SptP is a GTPase activating protein (GAP) for Rac1 and Cdc42 (ref. 4). In a precise matching of host cell function, SptP induces hydrolysis of GTP and shuts down the pathways controlled by the small GTPases that were activated by SopE and SopE2 (Fig. 1a).

The recently determined crystal structure of a SptP–Rac1 transition state complex reveals that the GAP domain of this effector, although possessing a new GAP fold, closely mimics host GAP enzymes⁵. Notably, this mimicry occurs through a combination of structural elements, some of which mirror precisely the chemical groups and interactions from host proteins, whereas others use similar amino acids in new contexts.

Like host-cell GAPs^{25–27}, SptP extensively interacts with the regulatory Switch I and II regions of the GTPase, contacting similar residues—especially catalytic residues—but doing so in a very different manner from host enzymes. Despite the different molecular tentacles extended by SptP relative to host proteins, the Switch I and II regions of SptP-bound Rac1 adopt nearly identical conformations as those observed in host GAP–RhoGTPase transition state complexes. For example, SptP uses several side chains to constrain the catalytic Gln 61, a central functional element that positions a nucleophilic water molecule for attack on the γ -phosphate of GTP. Positioning Gln 61 in this way is considered crucial to the hydrolysis reaction²⁸. Therefore, SptP has evolved to mimic host GAPs to achieve the same goal, but through its own unique methods⁵ (Fig. 2).

A second instance of mimicry in this structure occurs with a donated catalytic residue from SptP. The small GTPases of the Ras/Rho family lack a crucial catalytic arginine for the GTP hydrolysis reaction. GAPs for the small GTPases invariably insert an arginine side chain, thereby completing the active site²⁸. Mutagenesis had identified a probable candidate for such a residue in SptP (Arg 209). In analogy to host GAPs, which contain such arginines on extended loops (arginine ‘finger loops’), it was proposed that this was the structural foundation for the GAP activity of SptP (ref. 4). Indeed, this arginine was identified as a candidate for mutagenesis as it appeared to reside (on the basis of sequence) in a flexible loop-like region. The crystal structure of SptP–Rac1, however, provides a surprise. Arg 209 is the catalytic residue, and is inserted into the active site of Rac1 in a manner that is nearly identical to functionally equivalent arginines from host enzymes. The surprise is that, unlike host GAPs, Arg 209 is not in a loop, but instead extends from an α -helix in SptP (ref. 5; Fig. 2). Thus, this foreign enzyme with a different tertiary structure still achieves a precise chemical mimicry of host proteins. A structure involving Rac1 in complex with ExoS, an SptP homologue from the pathogen *Pseudomonas aeruginosa*, substantiates these results²².

Invasin and receptor substrate mimicry

Yersinia pseudotuberculosis uses the envelope protein invasin to bind host-cell β 1 integrin surface receptors, thereby manipulating signal transduction pathways in the host and contributing to bacterial attachment and internalization (ref. 29; Fig. 1b). The potency of invasin is such that it will out-compete natural host substrates for β 1 integrin binding (for example, fibronectin)^{30,31}. No sequence similarity between invasin and host proteins can be detected, but the crystal structure of invasin determined by Bjorkman and colleagues reveals the effectiveness of convergent evolution in producing a virulence factor that mimics host activities³. In this case, what is mimicked is the integrin-binding surface of fibronectin.

Both invasin and fibronectin appear in their respective crystal structures as long, rod-like molecules composed of repeated domains^{3,32}. Capping the rods in each case are specialized domains with the integrin-binding surface. At first glance, the tertiary structures of the integrin-binding domains of invasin and fibronectin seem completely unrelated. However, on examination of their integrin-binding surfaces, it becomes apparent that in spite of

Table 1 Examples of potential bacterial mimics

Virulence factor	Pathogen	Biological function	Biochemical activity
Horizontal acquisition*			
SptP (PTP domain) ⁵	<i>Salmonella</i> spp.	Reverses effects induced by bacterial internalization	Protein tyrosine phosphatase
YopH (ref. 4)	<i>Yersinia</i> spp.	Disrupts focal adhesions, paralyzing macrophages	Protein tyrosine phosphatase
YpkA (ref. 6)	<i>Yersinia</i> spp.	Disrupts the actin cytoskeleton	Serine/threonine kinase
YopJ (ref. 18)	<i>Yersinia</i> spp.	Inhibits MAP kinase and NF κ B signalling pathways	Cysteine protease
SopB (ref. 7)	<i>Salmonella</i> spp.	Promotes bacterial internalization, stimulates Cl [–] secretion, involved in enteropathogenicity	Inositol phosphate phosphatase
Convergent evolution†			
SptP (GAP domain) ⁵	<i>Salmonella</i> spp.	Reverses the host-cell actin cytoskeletal changes following bacterial internalization	GTPase activating protein for Cdc42 and Rac
YopE (ref. 21)	<i>Yersinia</i> spp.	Inhibits Rho GTPase family signalling, paralyzing macrophages	GTPase activating protein for Rho, Cdc42 and Rac
ExoS (ref. 22)	<i>Pseudomonas aeruginosa</i>	Inhibits Rho GTPase family signalling, paralyzing macrophages	GTPase activating protein for Rho, Cdc42 and Rac
SopE/E2 (refs 24, 39)	<i>Salmonella</i> spp.	Activates RhoGTPases, induces bacterial internalization	Guanine nucleotide exchange factor for Cdc42 and Rac
SipA (ref. 8)	<i>Salmonella</i> spp.	Promotes bacterial internalization	Reduces actin critical concentration, binds and stabilizes F-actin
SipC (ref. 40)	<i>Salmonella</i> spp.	Promotes bacterial internalization	Binds and nucleates actin
IpaA (ref. 41)	<i>Shigella</i> spp.	Promotes bacterial internalization	Binds vinculin and depolymerizes actin
IpaB (ref. 42)	<i>Shigella</i> spp.	Induces programmed cell death in macrophages	Binds and activates procaspase-1
ActA (ref. 34)	<i>Listeria monocytogenes</i>	Mediates intracellular movement and formation of actin tails	Promotes actin nucleation
IcsA/VirG (ref. 43)	<i>Shigella</i> spp.	Mediates intracellular movement and formation of actin tails	Promotes actin nucleation
Tir (ref. 44)	<i>Escherichia coli</i>	Promotes formation of actin pedestals	Binds Nck
Internalin (ref. 35)	<i>Listeria monocytogenes</i>	Promotes bacterial entry into host cells	Binds gC1q-R
InlB (ref. 36)	<i>Listeria monocytogenes</i>	Promotes bacterial entry into host cells	Binds the Met receptor tyrosine kinase
VacA (ref. 33)	<i>Helicobacter pylori</i>	Alters vesicular trafficking	Unknown

* Horizontal acquisition suggested by sequence and/or structural homology.

† Convergent evaluation suggested by lack of sequence and/or structural similarity.

the different protein architectures that scaffold the amino acids forming the binding site, the functional aspects of the molecular surfaces of invasin and fibronectin are similar (Fig. 3). At this level, invasin is an excellent mimic of the integrin-binding surface of fibronectin³².

Like SptP, invasin is a mixture of divergent elements and profound biochemical similarities when compared to the functional homologue of its host. For example, although fibronectin presents a binding surface that contains an extensive cleft absent in invasin, three residues that are required for integrin binding (two aspartic acids and an arginine) are located in nearly identical positions spanning the breadth of the extensive binding surface (Fig. 3). The fact that two completely different protein structures could nevertheless present the same residues in the same positions for binding attests to the power of convergent evolution. Invasin is therefore an

excellent mimic of its eukaryotic cell counterpart, an ability that provides *Yersinia* with access into host cells.

Evolutionary dynamics and the pathways to mimicry

From the examples discussed, one can consider the different pathways to mimicry. Bacterial virulence factors such as SptP and YopH, which must independently embody the full activity of this family of proteins, are likely to have been acquired horizontally (perhaps from an eukaryotic host). The convergent evolution of the necessary three-dimensional scaffolding and active site (quite elaborate in existing enzymes) would perhaps constitute a very long process of evolution for their acquisition.

In contrast, virulence factors that display their mimicry through molecular surfaces that mimic host protein surfaces are more likely to have been obtained by convergent evolution. Unlike the enzyme

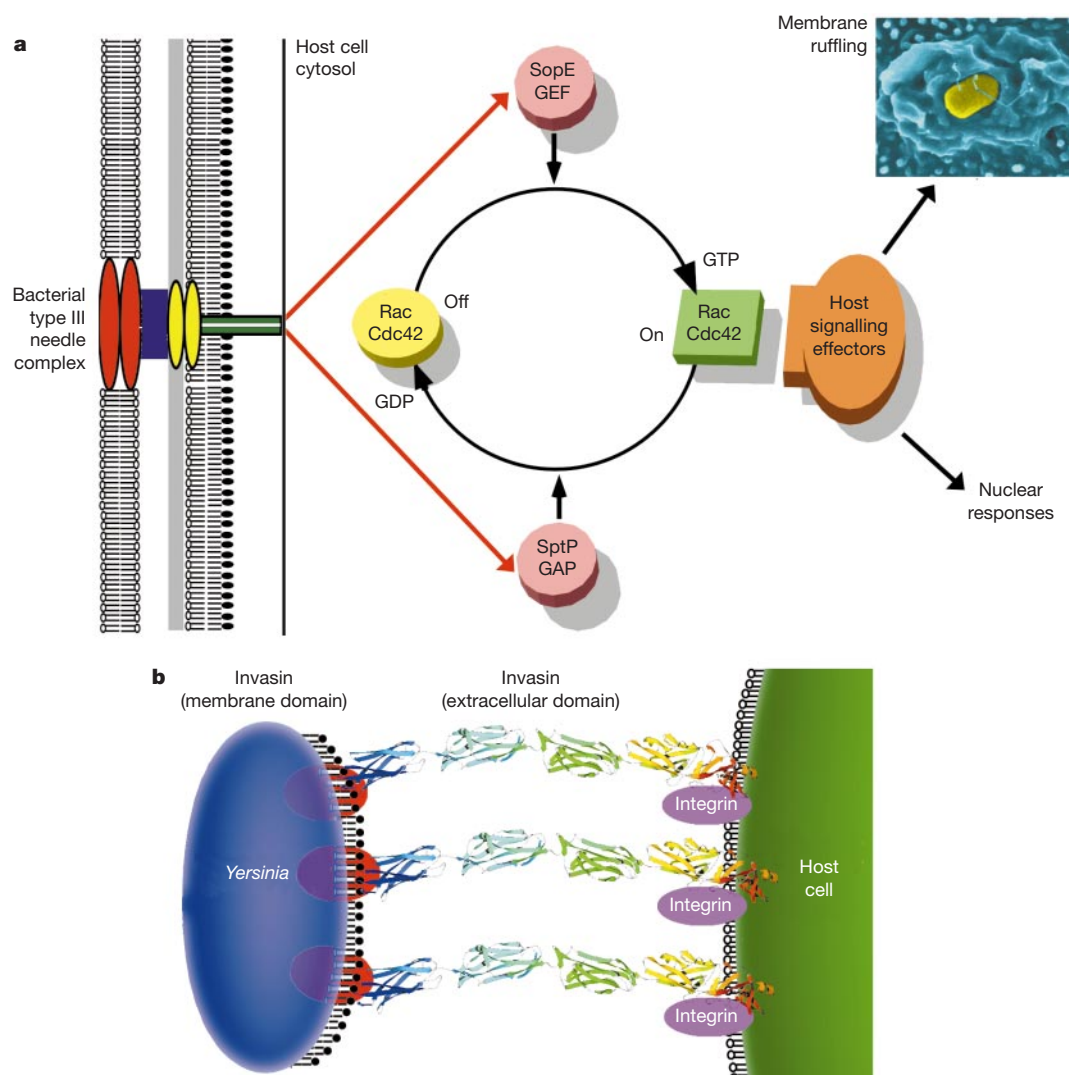


Figure 1 Host mimicry in the interaction of pathogenic bacteria with host cells. **a**, Interaction of *Salmonella* with host cells—striking a balance through mimicry. On contact with the host cell membrane, *Salmonella* uses a type III secretion system and its needle complex to deliver several bacterial effector proteins into the cell. These include SopE, which is an exchange factor for Rac and Cdc42, and SptP, a GAP for the same GTPases. SopE induces pronounced rearrangements of the actin cytoskeleton, membrane ruffling and nuclear responses by activating the Rho family GTPases Cdc42 and Rac1. The cytoskeletal alterations lead to the internalization of *Salmonella* and the nuclear responses result in pro-inflammatory cytokine production. After bacterial internalization, the bacterial effector SptP downregulates signalling from these GTPases, thereby shutting down the signals that promote these cellular responses. This contributes to the recovery of the host cell and helps to protect the cell from potential cytotoxic effects resulting from excessive Rho GTPase signalling. GEF, guanine nucleotide exchange factor. **b**, Invasin mediates *Yersinia* internalization by mimicking host integrin ligands. Invasin, with an outer-membrane-anchoring domain (red) and an integrin-binding extracellular domain (ribbon diagram coloured from blue at the N terminus through to red at the C terminus), binds the $\beta 1$ integrin receptor with high affinity, triggering signalling events that lead to bacterial internalization.

mimics, these bacterial proteins do not exert their functions through free-standing catalytic activities but rather by binding host proteins to modulate their activities. The integrin-binding domain of invasins and the GAP domain of SptP are two prominent examples of this pathway to mimicry. Through the precise arrangement of important residues on its surface, invasins mimics fibronectin's high-affinity binding to $\beta 1$ integrin receptors. For SptP, relatively modest architectural alterations of elements already available to the pathogen (for example, the ubiquitous four-helix bundle fold) may have been required to shape an appropriate interface and catalytic residue to evolve its GAP activity. Both of these examples should be qualified with the caveat that it is always possible that future studies may reveal structural homologues to these factors in host cells. For the time being, however, we may cautiously conclude that convergent evolution is the driving force behind the acquisition of these molecules.

It is probable that there are many other bacterial proteins that function as host mimics and are yet to be identified. Indeed, the intimate functional interface between the host and pathogen that is generated through type III secretion has produced a 'list of mimic suspects' among the translocated effectors that function within host cells, but show little if no sequence similarity to host proteins (Table 1). Although proteins of dissimilar sequence may have a common

structure, it is probable that at least some of these effectors may represent convergently evolved mimics, and not just examples of structural similarity concealed by low sequence homology. One may expect, therefore, that the large family of Gram-negative bacteria that use type III secretion systems will add to this collection of host mimics obtained from both the convergent and horizontal evolutionary pathways.

Other potential host mimics include the VacA protein of *Helicobacter pylori*, which on entry into host cells alters vesicular transport through mechanisms that are still unclear³³. VacA has no obvious homology to other proteins. Another example is the protein internalin. The internalin family of leucine-rich-repeat (LRR)-containing proteins is encoded by the bacterial pathogen *Listeria*³⁴. Two members of this family, internalin and InlB, mediate *Listeria* internalization into different host cells by interacting with specific surface receptors: E-cadherin in the case of internalin and the Met or gC1q-R protein in the case of InlB^{35,36}. LRRs are protein-protein interaction modules attached to diverse molecules with disparate functions, both in bacteria and eukaryotic cells³⁷. The crystal structure of InlB, determined by Ghosh and colleagues, raises some intriguing possibilities³⁸. The InlB LRRs are structured like host homologues (although with some interesting twists), present a binding surface like eukaryotic LRRs, and one of the repeats is

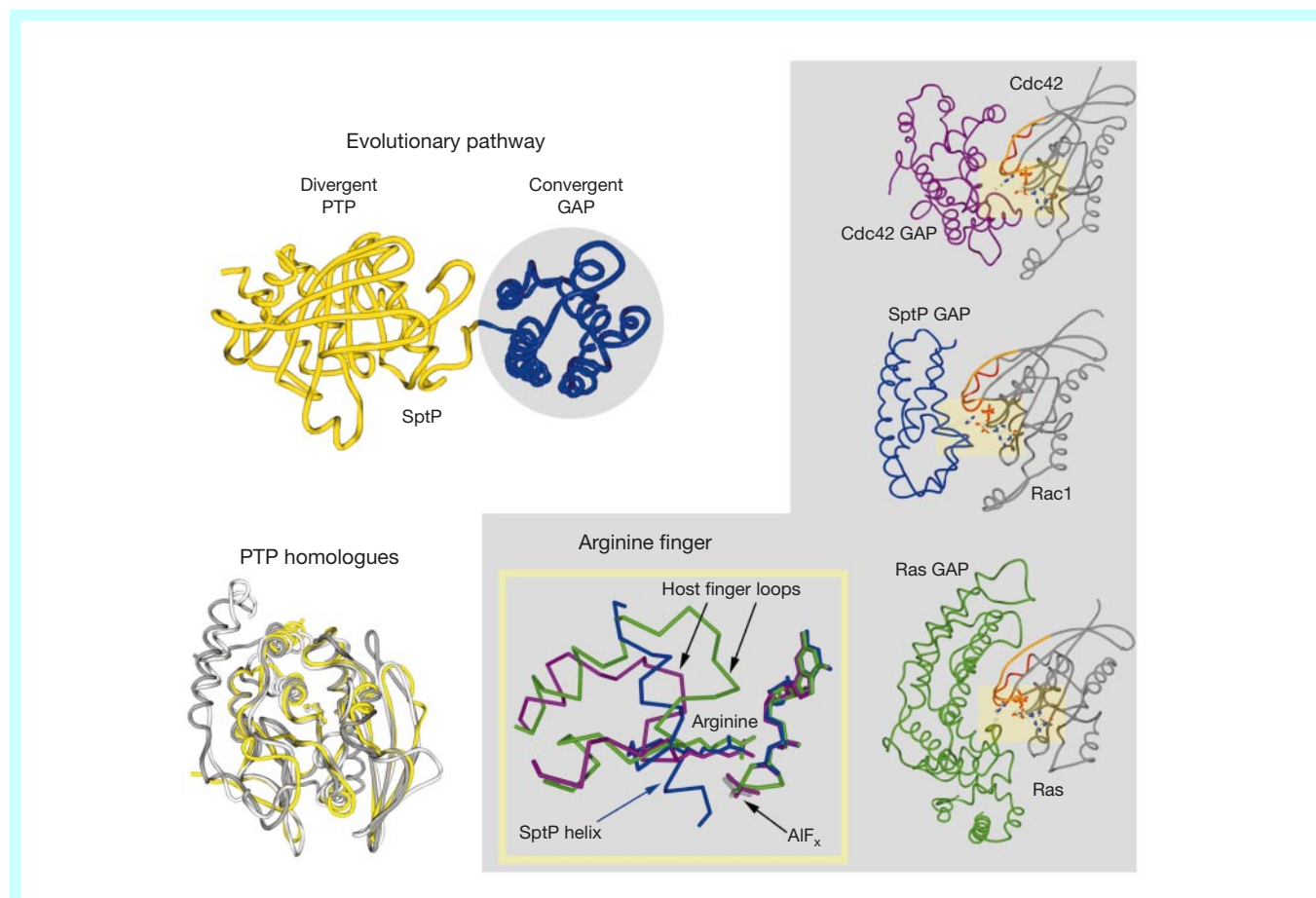


Figure 2 Construction of a virulence factor by different evolutionary pathways. The *Salmonella* virulence factor SptP is composed of two independent modules: a four-helix bundle, Rho GTPase activating (GAP) domain (blue) and a canonical tyrosine phosphatase domain (yellow). The crystal structure of SptP (top left) suggests different evolutionary pathways for the acquisition of each of these domains. The tyrosine phosphatase domain of SptP shows extensive structural and mechanistic similarities to eukaryotic and bacterial homologues, as shown by its alignment with the PTP1B and YopH proteins. This suggests an evolutionary pathway involving horizontal acquisition. In contrast, the GAP domain exhibits a different fold from that of other eukaryotic homologues, which suggests that it is the product of convergent evolution. The structures of three transition-state complexes between small GTPases and their cognate GAPs are shown (grey background). The three GAPs engage the small G proteins in a similar fashion, primarily—and in the case of SptP, exclusively—by contacting the Switch I (orange), Switch II (red) and nucleotide regions of these proteins. A yellow box marks the active sites of these complexes with the nucleotide and catalytic arginine present in all known GAPs. A closer view of this region (yellow outline) illustrates that despite using a similar chemistry to the host factors, SptP (in blue) presents the arginine from a completely different protein architecture. AlF_x, aluminium fluoride.

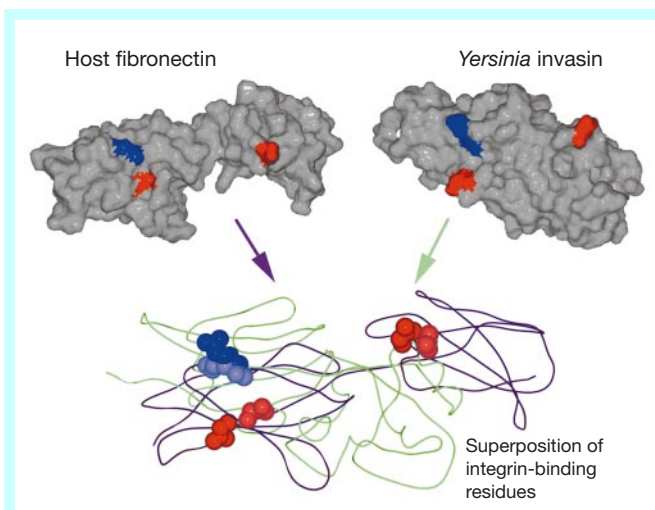


Figure 3 The integrin-binding region of invasin mimics the host integrin ligand, fibronectin. The molecular surfaces for these two domains are shown (top). The locations of the integrin-contacting aspartic acids (red) and arginine (blue) are shown. The proteins (invasin in green; fibronectin in purple) are aligned on the basis of minimizing the distance between these integrin-contacting residues (bottom). Despite the dissimilar tertiary folds and molecular surfaces, these residues align almost perfectly. This indicates that invasin has convergently evolved to be a precise host mimic.

thought to be involved in an intermolecular calcium-binding site with its receptor that may be important for function³⁸. It will be interesting to see whether these LRRs are an example of bacterial virulence factors shaped by convergent evolution of a receptor-binding domain (like invasin), or whether they constitute a case of a yet-unknown cellular-receptor-binding module that has been subverted for the pathogen's needs.

These studies have revealed several important points regarding bacterial virulence: (1) bacteria can have extremely subtle and sophisticated mechanisms for achieving their goals by means of fine-tuned and highly efficient biochemical processes; (2) molecular mimicry of host proteins is a powerful tool exploited by many bacterial pathogens in the process of host manipulation; and (3) convergent evolution contributes significantly to the dynamics of the evolutionary process. The insights to be gained from studies of bacterial virulence factors represent an opportunity for research into infectious agents, host cell biology and the evolution of pathogenesis. Structural studies in particular will be vital for a full appreciation of the mechanisms that these factors use to subvert the host. □

- Guan, K. & Dixon, J. E. Protein tyrosine phosphatase activity of an essential virulence determinant in *Yersinia*. *Science* **249**, 553–556 (1990).
- Hardt, W.-D., Chen, L.-M., Schuebel, K. E., Bustelo, X. R. & Galán, J. E. *Salmonella typhimurium* encodes an activator of Rho GTPases that induces membrane ruffling and nuclear responses in host cells. *Cell* **93**, 815–826 (1998).
- Hamburger, Z. A., Brown, M. S., Isberg, R. R. & Bjorkman, P. J. Crystal structure of invasin: a bacterial integrin-binding protein. *Science* **286**, 291–295 (1999).
- Fu, Y. & Galán, J. E. A *Salmonella* protein antagonizes Rac-1 and Cdc42 to mediate host-cell recovery after bacterial invasion. *Nature* **401**, 293–297 (1999).
- Stebbins, C. E. & Galán, J. E. Modulation of host signaling by a bacterial mimic. Structure of the *Salmonella* effector SptP bound to Rac1. *Mol. Cell* **6**, 1449–1460 (2000).
- Galyov, E. E., Hakansson, S., Forsberg, A. & Wolf-Watz, H. A secreted protein kinase of *Yersinia pseudotuberculosis* is an indispensable virulence determinant. *Nature* **361**, 730–732 (1993).
- Norris, F. A., Wilson, M. P., Wallis, T. S., Galyov, E. E. & Majerus, P. W. SopB, a protein required for virulence of *Salmonella dublin*, is an inositol phosphate phosphatase. *Proc. Natl Acad. Sci. USA* **95**, 14057–14059 (1998).
- Zhou, D., Mooseker, M. & Galán, J. E. Role of the *S. Typhimurium* actin-binding protein SipA in bacterial internalization. *Science* **283**, 2092–2095 (1999).
- Lerm, M., Schmidt, G. & Aktories, K. Bacterial protein toxins targeting rho GTPases. *FEMS Microbiol. Lett.* **188**, 1–6 (2000).
- Alouf, J. E. Bacterial protein toxins. An overview. *Methods Mol. Biol.* **145**, 1–26 (2000).
- Montecucco, C., Papini, E. & Schiavo, G. Bacterial protein toxins and cell vesicle trafficking. *Experientia* **52**, 1026–1032 (1996).
- Haag, F. & Koch-Nolte, F. Endogenous relatives of ADP-ribosylating bacterial toxins in mice and men:

- potential regulators of immune cell function. *J. Biol. Regul. Homeost. Agents* **12**, 53–62 (1998).
- Galán, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**, 322–328 (1999).
- Stuckey, J. A. *et al.* Crystal structure of *Yersinia* protein tyrosine phosphatase at 2.5 Å and the complex with tungstate. *Nature* **370**, 571–575 (1994).
- Kaniga, K., Uralil, J., Bliska, J. B. & Galán, J. E. A secreted tyrosine phosphatase with modular effector domains encoded by the bacterial pathogen *Salmonella typhimurium*. *Mol. Microbiol.* **21**, 633–641 (1996).
- Persson, C., Carballeira, N., Wolf-Watz, H. & Fallman, M. The PTPase YopH inhibits uptake of *Yersinia*, tyrosine phosphorylation of p130Cas and FAK, and the associated accumulation of these proteins in peripheral focal adhesions. *EMBO J.* **16**, 2307–2318 (1997).
- Black, D. S. & Bliska, J. B. Identification of p130Cas as a substrate of *Yersinia* YopH (Yop51), a bacterial protein tyrosine phosphatase that translocates into mammalian cells and targets focal adhesions. *EMBO J.* **16**, 2730–2744 (1997).
- Orth, K. *et al.* Disruption of signaling by the *Yersinia* effector YopJ, a ubiquitin-like protein protease. *Science* **290**, 1594–1597 (2000).
- Isberg, R. R. & Leong, J. M. Multiple beta 1 chain integrins are receptors for invasin, a protein that promotes bacterial penetration into mammalian cells. *Cell* **60**, 861–871 (1990).
- Goehring, U. M., Schmidt, G., Pederson, K. J., Aktories, K. & Barbieri, J. T. The N-terminal domain of *Pseudomonas aeruginosa* exoenzyme S is a GTPase-activating protein for Rho GTPases. *J. Biol. Chem.* **274**, 36369–36372 (1999).
- Von Pawel-Rammingen, U. *et al.* GAP activity of the *Yersinia* YopE cytotoxin specifically targets the rho pathway: a mechanism for disruption of actin microfilament structure. *Mol. Microbiol.* **36**, 737–748 (2000).
- Wurtele, M. *et al.* How the *Pseudomonas aeruginosa* ExoS toxin downregulates Rac. *Nature Struct. Biol.* **8**, 23–26 (2001).
- Galán, J. E. & Zhou, D. Striking a balance: modulation of the actin cytoskeleton by *Salmonella*. *Proc. Natl Acad. Sci. USA* **97**, 8754–8761 (2000).
- Stender, S. *et al.* Identification of SopE2 from *Salmonella typhimurium*, a conserved guanine nucleotide exchange factor for Cdc42 of the host cell. *Mol. Microbiol.* **36**, 1206–1221 (2000).
- Nassar, N., Hoffman, G. R., Mannor, D., Clardy, J. C. & Cerione, R. A. Structures of Cdc42 bound to the active and catalytically compromised forms of Cdc42GAP. *Nature Struct. Biol.* **5**, 1047–1052 (1998).
- Rittinger, K., Walker, P. A., Eccleston, J. F., Smerdon, S. J. & Gamblin, S. J. Structure at 1.65 Å of RhoA and its GTPase-activating protein in complex with a transition-state analogue. *Nature* **389**, 758–762 (1997).
- Scheffzek, K. *et al.* The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333–338 (1997).
- Scheffzek, K., Ahmadian, M. R. & Wittinghofer, A. GTPase-activating proteins: helping hands to complement an active site. *Trends Biochem. Sci.* **23**, 7257–7262 (1998).
- Isberg, R. R., Hamburger, Z. & Dersch, P. Signaling and invasin-promoted uptake via integrin receptors. *Microbes Infect.* **2**, 793–801 (2000).
- Tran Van Nhieu, G. & Isberg, R. R. The *Yersinia pseudotuberculosis* invasin protein and human fibronectin bind to mutually exclusive sites on the alpha 5 beta 1 integrin receptor. *J. Biol. Chem.* **266**, 24367–24375 (1991).
- Tran Van Nhieu, G. & Isberg, R. R. Bacterial internalization mediated by beta 1 chain integrins is determined by ligand affinity and receptor density. *EMBO J.* **12**, 1887–1895 (1993).
- Leahy, D. J., Aukhil, I. & Erickson, H. P. 2.0 Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell* **84**, 155–164 (1996).
- Reyrat, J. M. *et al.* Towards deciphering the *Helicobacter pylori* cytotoxin. *Mol. Microbiol.* **34**, 197–204 (1999).
- Cossart, P. & Lecuit, M. Interactions of *Listeria monocytogenes* with mammalian cells during entry and actin-based movement: bacterial factors, cellular ligands and signaling. *EMBO J.* **17**, 3797–3806 (1998).
- Braun, L., Ghebrehewet, B. & Cossart, P. gC1q-R/p32, a C1q-binding protein, is a receptor for the InlB invasion protein of *Listeria monocytogenes*. *EMBO J.* **19**, 1458–1466 (2000).
- Shen, Y., Naujokas, M., Park, M. & Ireton, K. InlB-dependent internalization of *Listeria* is mediated by the met receptor tyrosine kinase. *Cell* **103**, 501–510 (2000).
- Kobe, B. & Deisenhofer, J. The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci.* **19**, 415–421 (1994).
- Marino, M., Braun, L., Cossart, P. & Ghosh, P. Structure of the InlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen *Listeria monocytogenes*. *Mol. Cell* **4**, 1063–1072 (1999).
- Hardt, W.-D., Urlaub, H. & Galán, J. E. A target of the centisome 63 type III protein secretion system of *Salmonella typhimurium* is encoded by a cryptic bacteriophage. *Proc. Natl Acad. Sci. USA* **95**, 2574–2579 (1998).
- Hayward, R. D. & Koronakis, V. Direct nucleation and bundling of actin by the SipC protein of invasive *Salmonella*. *EMBO J.* **18**, 4926–4934 (1999).
- Tran Van Nhieu, G., Ben-Ze'ev, A. & Sansonetti, P. J. Modulation of bacterial entry into epithelial cells by association between vinculin and the Shigella IpaA invasin. *EMBO J.* **16**, 2717–2729 (1997).
- Chen, Y., Smith, M. R., Thirumalai, K. & Zychlinsky, A. A bacterial invasin induces macrophage apoptosis by binding directly to ICE. *EMBO J.* **15**, 3853–3860 (1996).
- Bourdet-Sicard, R., Egile, C., Sansonetti, P. J. & Tran Van Nhieu, G. Diversion of cytoskeletal processes by *Shigella* during invasion of epithelial cells. *Microbes Infect.* **2**, 813–819 (2000).
- Goosney, D. L., Gruenheid, S. & Finlay, B. B. Gut feelings: enteropathogenic *E. coli* (EPEC) interactions with the host. *Annu. Rev. Cell Dev. Biol.* **16**, 173–189 (2000).

Acknowledgements

C.E.S. was supported by a fellowship of the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation. This work was supported by Public Health Services grants to J.E.G.

Correspondence and requests for materials should be addressed to J.E.G. (e-mail: jorge.galan@yale.edu). Coordinates have been deposited in the Protein DataBank under codes 1G4U and 1G4W for the SptP-Rac1 heterodimer and SptP monomer, respectively.

Deficiency of molecular hydrogen in the disk of β Pictoris

A. Lecavelier des Etangs*, A. Vidal-Madjar*, A. Roberge†, P. D. Feldman‡, M. Deleuil‡, M. André†, W. P. Blair†, J.-C. Bouret‡, J.-M. Désert*, R. Ferlet*, S. Friedman†, G. Hébrard*, M. Lemoine* & H. W. Moos†

* Institut d'Astrophysique de Paris, CNRS, 98 bis bld Arago, F-75014 Paris, France

† Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA

‡ Laboratoire d'Astrophysique de Marseille, BP 8, F-13376 Marseille Cedex 12, France

Molecular hydrogen (H_2) is by far the most abundant material from which stars, protoplanetary disks and giant planets form, but it is difficult to detect directly. Infrared emission lines from H_2 have recently been reported¹ towards β Pictoris, a star harbouring a young planetary system². This star is surrounded by a dusty 'debris disk' that is continuously replenished either by collisions between asteroidal objects³ or by evaporation of ices on Chiron-like objects⁴. A gaseous disk has also been inferred from absorption lines in the stellar spectrum^{5–8}. Here we present the far-ultraviolet spectrum of β Pictoris, in which H_2 absorption lines are not seen. This allows us to set a very low upper limit on the column density of H_2 : $N(H_2) \leq 10^{18} \text{ cm}^{-2}$. This non-detection is puzzling when compared to the quantity of H_2 inferred from the infrared observations, but it does show that H_2 is not in the disk on the direct line of sight. Carbon monoxide (CO) has been seen in absorption against the star^{8–10}, yielding a ratio of $CO/H_2 > 6 \times 10^{-4}$. As CO would be destroyed under ambient conditions in about 200 years (refs 9, 11), our result demonstrates that the CO in the disk arises from evaporation of planetesimals.

The age of β Pic (about 20 Myr; ref. 12) indicates that the planetary system around this star is likely to be in the late stage of formation—the clearing of the remaining planetesimals. An extensive spectroscopic survey has revealed very rapid spectral variations, interpreted as hundreds of star-grazing comets passing in front of the star^{13,14}. The gas component also shows the presence of CO, seen in ultraviolet (UV) absorption lines^{8–10}.

Observation of far-UV electronic transitions is the most powerful tool for detection of cold molecular hydrogen. Although very abundant, this symmetric molecule does not radiate at radio wavelengths, and thus remained undetected until its absorption signatures from Werner and Lyman series were detected by space

missions, such as the Copernicus satellite in the early 1970s^{15–17}. As absorption lines of CO were easily detected towards β Pic, it was expected that H_2 would be detected by the Far Ultraviolet Spectroscopic Explorer (FUSE)¹⁸ satellite. Moreover, a large quantity of H_2 has recently been inferred to exist in the β Pic system, on the basis of detection by the ISO satellite of its infrared emission lines¹. If this H_2 is co-spatial with the dust, it should be readily detectable by far-UV spectroscopy of the star seen through its edge-on disk.

FUSE observations of β Pic were made on 18 March 2000 with a test exposure of 5,800 s, and on 1 and 3 March 2001 for a total time of about 28,800 s. This star is very faint in the far-UV, and hence its continuum is almost undetectable at the wavelengths of the H_2 transitions (Fig. 1): in the 1,108-Å region where we could expect to see the H_2 (0–0) band in absorption, the signal-to-noise ratio is less than 1 per 0.1 Å. However, the clear presence of the stellar O VI emission doublet at 1,031.9 and 1,037.6 Å, and the faint emission from C II at 1,036.3 Å, provide the opportunity to strongly constrain the H_2 content along the line of sight. With rest wavelengths of 1,036.55 Å ($J = 0$), 1,037.15 Å and 1,038.16 Å ($J = 1$), and 1,038.69 Å ($J = 2$), the H_2 (5–0) lines are exactly superimposed on the weaker line of the O VI doublet. The H_2 lines at $J = 3$ (1,031.20 Å) and $J = 4$ (1,032.35 Å) are superimposed on the stronger line of the doublet.

The O VI emission lines that we detect cannot originate in the interstellar medium, nor can they be scattered solar light. Indeed, the interstellar diffuse emission is 100 times fainter and, furthermore, simultaneous spectra acquired with the MDRS (medium resolution) aperture of FUSE, pointed about 200 arcsec away from β Pic, do not show any emission other than airglow lines. These O VI lines thus originate from β Pic¹⁹, and allow us to probe the line of sight towards the star through the edge-on circumstellar disk.

If H_2 were present in substantial quantity in the line of sight, the O VI line at 1,038 Å would be absorbed, and hence would appear abnormally weak—less than half the intensity of the other O VI line at 1,032 Å (Fig. 2). For instance, in the spectrum of the Herbig Ae star AB Aurigae, the O VI line at 1,038 Å is almost totally absorbed and a corresponding H_2 column density of $7 \times 10^{19} \text{ cm}^{-2}$ has been derived²⁰. In β Pic, emission is seen in both lines of the O VI doublet with an integrated flux ratio close to the expected value of 2. This implies that H_2 is not present in a quantity sufficient to significantly absorb the O VI emission around 1,038 Å. We have detected more than 6,000 photons in the O VI line, compared to 3,000 photons due to the background noise; H_2 column density above a few times 10^{20} cm^{-2} is thus excluded at the $\sim 45\sigma$ level.

A better upper limit on the amount of H_2 can be estimated by assuming the intrinsic width of the H_2 lines to be very small. The absence of wide absorption lines thus gives a conservative upper

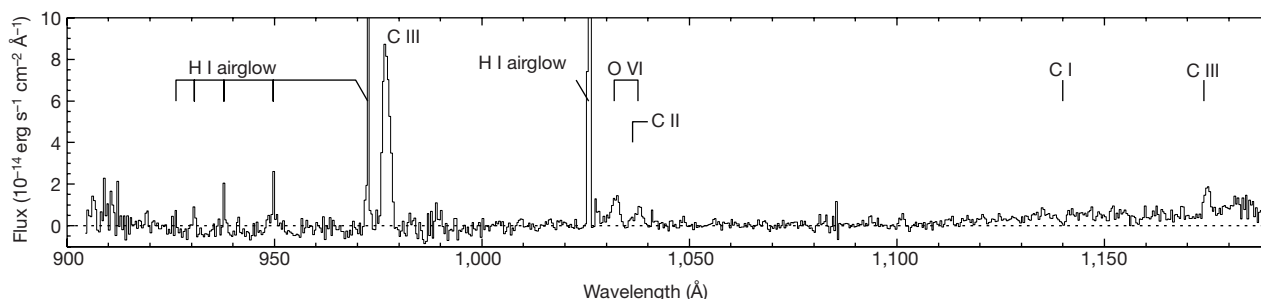


Figure 1 Far ultraviolet spectrum of β Pic from 905 to 1,188 Å, as obtained by the FUSE satellite. The target was acquired in the LWRS (low resolution) slit of $30'' \times 30''$ after a blind offset from a reference star. The data have been processed with the 'calfuse' pipeline software, version 1.8.7. The atmospheric H I airglow lines of the Lyman series are detected from Ly β (1,025.7 Å) to Ly η (926.2 Å). We also detect the stellar continuum

down to $\sim 1,100$ Å. This continuum is well reproduced by a Kurucz model with an effective temperature of 8,200 K, a turbulent velocity of 2 km s^{-1} , and surface gravity $\log g = 4.25$. C I photospheric lines predicted by the stellar model are seen near 1,140 Å. From the position of a few stellar photospheric lines, the uncertainty in the wavelength calibration is less than 40 km s^{-1} . C II, C III and O VI are also detected in emission.

limit on the column density of H_2 , which corresponds to the limit at which the 'wings' of saturated absorption lines in the lorentzian part of the profile would start to become detectable.

We may have marginally detected an absorption line of H_2 at $J = 0$. We infer a corresponding column density $N_{J=0}$ of $8 \times 10^{17} \text{ cm}^{-2}$, assuming that the emission line is smooth in this wavelength region. However, because the line is located between the C II and O VI emission lines, this 'detection' could be an artefact arising from the lack of emission precisely at this wavelength. In any case, the absence of a wide absorption line allows us to constrain the column density of H_2 at $J = 0$. A column density $N_{J=0} = 3 \times 10^{18} \text{ cm}^{-2}$ increases the χ^2 of the fit to the data by 4 in comparison to the fit of the emission lines without H_2 . We conclude that the column density of H_2 at $J = 0$ is below $3 \times 10^{18} \text{ cm}^{-2}$ at the 95% confidence level.

The column densities of H_2 at $J = 1$ to $J = 4$ are better constrained, because they are well superimposed on the O VI lines. We find the following upper limits at the 95% confidence level: $N_{J=1} \leq 5$

$\times 10^{17} \text{ cm}^{-2}$, $N_{J=2} \leq 8 \times 10^{17} \text{ cm}^{-2}$, $N_{J=3} \leq 4 \times 10^{17} \text{ cm}^{-2}$ and $N_{J=4} \leq 1 \times 10^{17} \text{ cm}^{-2}$. The upper limit for $N_{J=0}$ is slightly less well constrained. However, the probability that $N_{J=0}$ is larger than $N_{J=1}$ is low; this would require a temperature lower than 20 K at thermodynamic equilibrium. With the assumption that $N_{J=0} \leq N_{J=1}$, we obtain an upper limit for the total H_2 column density toward β Pic: $N(H_2) \leq 10^{18} \text{ cm}^{-2}$ at the 95% confidence level (Fig. 2).

This upper limit on the H_2 column density obtained from the FUSE spectrum is very low, which has several consequences. First, the CO detected in UV spectra of β Pic (obtained by the Hubble Space Telescope) has a column density $N(\text{CO})$ of $6 \times 10^{14} \text{ cm}^{-2}$ (ref. 8). Without any geometrical assumptions, we find $N(\text{CO})/N(H_2) > 6 \times 10^{-4}$. This ratio towards β Pic is two orders of magnitude higher than the value for interstellar translucent clouds having a similarly low column density of CO (ref. 21). This proves that the CO is not interstellar, but is instead located in the disk. However, the published modelling of the β Pic CO abundance—which assumes chemical equilibrium within a dense molecular circumstellar disk—predicts a H_2 column density of the order of 10^{22} cm^{-2} (ref. 22), in disagreement with the present observation. Such a high column density is required to shield CO from dissociation by interstellar UV photons, and thus to allow the formation of CO. The present result shows that the gaseous disk has a low density, and that CO cannot form by the same chemical reaction that produces interstellar CO. On the contrary, at the estimated low column density of H_2 , photodissociation dominates, destroying CO on a timescale of about 200 years (refs 9, 11). This implies that CO must be re-supplied to the circumstellar disk by release from a pre-existing reservoir. Two processes have been proposed: either evaporation of star-grazing comets, or a slow evaporation of planetesimals on moderately eccentric orbits at several tens of astronomical units (AU) from the star, as observed for Chiron in the Solar System²³. The evaporation naturally explains the large CO/ H_2 ratio: CO is trapped in the ices sublimated from the evaporating bodies, whereas H_2 cannot be trapped in this way.

The present H_2 upper limit is well below the recently reported H_2 detection from ISO observations using the SWS spectrometer¹. From the two emission lines of H_2 at 17 and 28 μm , a total H_2 mass of 0.17 Jupiter masses was inferred within the ISO beam ($280 \text{ AU} \times 540 \text{ AU}$ and $400 \text{ AU} \times 540 \text{ AU}$ for the two lines, respectively). If this H_2 were distributed like the dust in the edge-on disk, a column density of 5×10^{20} to $5 \times 10^{22} \text{ cm}^{-2}$ should have been detected in absorption, depending on the geometry and inclination of the system. In other words, the present upper limit on the H_2 column density constrains the total mass in the disk to be below $\sim 3 \times 10^{-4}$ Jupiter masses. This discrepancy of three orders of magnitude needs to be explained.

ISO has detected H_2 only in the high rotational levels $J = 2$ and $J = 3$, which contain less than 2% of the total mass, at a temperature of 109 K determined by the ratio of the two emission lines. FUSE gives access to the $J = 0$ and $J = 1$ levels as well, where most of the H_2 is found, at the typical temperatures of the interstellar medium or circumstellar disks. The present upper limit for H_2 ($J = 2$) implies that the mass of H_2 ($J = 2$) in the disk is more than ten times smaller than the mass of H_2 ($J = 2$) seen in the infrared.

We therefore conclude that the H_2 detected by ISO cannot be uniformly distributed within the disk of β Pic (the opening angle of the dust disk is larger than its inclination). Nor can it be distributed in a single large cloud in front of β Pic: 0.17 Jupiter masses distributed homogeneously across the full ISO beam would yield a H_2 column density of the order of $2 \times 10^{21} \text{ cm}^{-2}$. Any such cloud in the foreground of β Pic would have been readily detected. The possibility that ISO detected an interstellar molecular cloud beyond β Pic is also excluded, because the non-detection of CO radio emission in a similar beam would imply a very abnormal CO/ H_2 ratio²⁴. Finally, we note that, because the O VI emission lines are more than 2 Å wide, only H_2 gas Doppler-shifted by more than

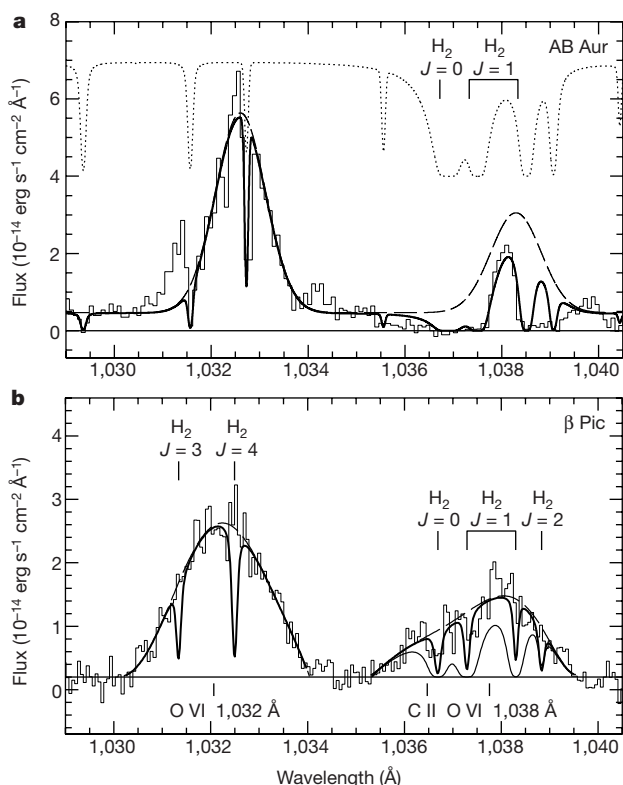


Figure 2 Plot of the observed O VI emission doublet and H_2 absorption lines. **a**, For the Herbig star AB Aur²⁰, the O VI line at 1,032 Å is well fitted by a gaussian (dashed line). The O VI emission line at 1,038 Å is almost completely absorbed by a H_2 column density of $N(H_2) = 7 \times 10^{19} \text{ cm}^{-2}$. The normalized H_2 model is shown with a dotted line. Only a small window between the two $J = 1$ lines of H_2 allows a fraction of the emission to be seen (solid line). **b**, For β Pic, the wide O VI and C II emission lines are fitted with a polynomial (dashed line). The flux ratio of the two O VI lines is close to the theoretical value of 2, providing a direct indication of the low column density of H_2 . The thick solid line shows the expected spectrum for a H_2 column density of $N(H_2) = 10^{18} \text{ cm}^{-2}$ in each J -level. Except for $J = 0$, such a high column density is excluded at more than the 99% confidence level. The thin solid line shows the modelled spectrum if H_2 were present with a total column density of $5 \times 10^{19} \text{ cm}^{-2}$ and a temperature of 109 K, as given by the ISO observations of the 17- and 28- μm emission lines¹. At this temperature, 34% of the H_2 molecules are at $J = 0$, whereas 64% are at $J = 1$. The total column density of $5 \times 10^{19} \text{ cm}^{-2}$ is actually only one-tenth of the lowest possible H_2 column density indicated by the ISO observations, if the H_2 were co-spatial with the dust. Such a model is rejected at more than $\sim 20\sigma$. If the H_2 detected by ISO were uniformly distributed in the disk, the O VI emission line at 1,038 Å would be completely absorbed and rendered undetectable in the present observation.

600 km s⁻¹ could escape detection; this is unlikely, as 600 km s⁻¹ is greater than the escape velocity. Consequently, neither *J*-level population (*J* = 2 is not detected) nor Doppler shift can explain the discrepancy. If the H₂ seen in the infrared were located in the disk, it would have been detected in our FUSE observation.

A solution to this problem is to consider that H₂ is not distributed widely throughout the disk. If the ISO detection is confirmed by further observations, we suggest that the H₂ could be confined into individual clouds, none of which intersected the small volume of the β Pic line of sight at the time of our FUSE observations. Such H₂ clouds would easily escape detection in absorption, but would be seen in emission. The nature of these clouds remains unclear; however, we speculate that they might be remnants of gaseous planet embryos which captured a significant amount of protoplanetary material before its recent dissipation. The stability, number and density of these putative clouds need to be analysed. Given that 0.17 Jupiter masses is a huge amount of gas, these clouds should have large physical sizes and/or be relatively numerous. The discrepancy between the ISO detection and the present negative FUSE result remains a challenging issue.

Using ISO, even larger amounts of H₂ have been reported¹ for other similar, albeit younger, circumstellar disks. Again the nature and the geometrical distribution of this H₂ remain to be determined. The present result shows that, in the circumstellar disks, dust is not a good indicator of the H₂ distribution; it also shows that CO may not generally be as depleted in the disks as may be concluded simply from the observed ratio of CO and H₂ emission. On the contrary, CO could be over-abundant compared to the standard CO/H₂ ratio; this provides a clue to the evaporation activity related to large-scale motions of the remnant planetesimals being cleared from the young planetary disk. □

Received 14 March; accepted 12 June 2001.

1. Thi, W. F. *et al.* Substantial reservoirs of molecular hydrogen in the debris disks around young stars. *Nature* **409**, 60–63 (2001).
2. Vidal-Madjar, A., Lecavelier des Etangs, A. & Ferlet, R. Beta Pictoris, a young planetary system? A review. *Planet. Space Sci.* **46**, 629–648 (1998).
3. Backman, D. E. & Paresce, F. Main-sequence stars with circumstellar solid material—the Vega phenomenon. *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1253–1304 (Univ. Arizona Press, 1993).
4. Lecavelier des Etangs, A., Vidal-Madjar, A. & Ferlet, R. Dust distribution in disks supplied by small bodies: is the Beta Pictoris disk a gigantic multi-cometary tail? *Astron. Astrophys.* **307**, 542–550 (1996).
5. Hobbs, L. M., Vidal-Madjar, A., Ferlet, R., Albert, C. E. & Gry, C. The gaseous component of the disk around Beta Pictoris. *Astrophys. J.* **293**, L29–L33 (1985).
6. Vidal-Madjar, A., Ferlet, R., Hobbs, L. M., Gry, C. & Albert, C. E. The circumstellar gas cloud around Beta Pictoris. II. *Astron. Astrophys.* **167**, 325–332 (1986).
7. Lagrange, A.-M. *et al.* The Beta Pictoris circumstellar disk. XXIV. Clues to the origin of the stable gas. *Astron. Astrophys.* **330**, 1091–1108 (1998).
8. Roberge, A. *et al.* High-resolution Hubble Space Telescope STIS spectra of CI and CO in the Beta Pictoris circumstellar disk. *Astrophys. J.* **538**, 904–910 (2000).
9. Vidal-Madjar, A. *et al.* HST-GHRS observations of Beta Pictoris: additional evidence for infalling comets. *Astron. Astrophys.* **290**, 245–258 (1994).
10. Jolly, A. *et al.* HST-GHRS observations of CO and CI in the Beta Pictoris circumstellar disk. *Astron. Astrophys.* **329**, 1028–1034 (1998).
11. van Dishoeck, E. F. & Black, J. H. The photodissociation and chemistry of interstellar CO. *Astrophys. J.* **334**, 771–802 (1988).
12. Barrado y Navascués, D., Stauffer, J. R., Song, I. & Caillault, J.-P. The age of Beta Pictoris. *Astrophys. J.* **520**, L123–L126 (1999).
13. Ferlet, R., Vidal-Madjar, A. & Hobbs, L. M. The Beta Pictoris circumstellar disk. V. Time variations of the CA II-K line. *Astron. Astrophys.* **185**, 267–270 (1987).
14. Beust, H., Vidal-Madjar, A., Ferlet, R. & Lagrange-Henri, A. M. The Beta Pictoris circumstellar disk. X-Numerical simulations of infalling evaporating bodies. *Astron. Astrophys.* **236**, 202–216 (1990).
15. Carruthers, G. R. Rocket observation of interstellar molecular hydrogen. *Astrophys. J.* **161**, L81–L85 (1970).
16. Spitzer, L. *et al.* Spectrophotometric results from the Copernicus satellite. IV. Molecular hydrogen in interstellar space. *Astrophys. J.* **181**, L116–L121 (1973).
17. Shull, J. M. & Beckwith, S. Interstellar molecular hydrogen. *Annu. Rev. Astron. Astrophys.* **20**, 163–190 (1982).
18. Moos, H. W. *et al.* Overview of the Far Ultraviolet Spectroscopic Explorer Mission. *Astrophys. J.* **538**, L1–L6 (2000).
19. Deleuil, M. *et al.* Is Beta Pictoris an active star? *Astrophys. J.* (in the press).
20. Roberge, A. *et al.* FUSE and HST STIS observations of hot and cold gas in the AB Aurigae system. *Astrophys. J.* **551**, L97–L100 (2001).
21. Magnani, L., Onello, J. S., Adams, N. G., Hartmann, D. & Thaddeus, P. The variation of the CO to H₂ conversion factor in two translucent clouds. *Astrophys. J.* **504**, 290–299 (1998).

22. Kamp, I. & Bertoldi, F. CO in the circumstellar disks of Vega and Beta Pictoris. *Astron. Astrophys.* **353**, 276–286 (2000).
23. Lecavelier des Etangs, A. Circumstellar disks and outer planet formation. *Planets Outside the Solar System: Theory and Observations* (eds Mariotti, J.-M. & Alloin, D.) 95–103 (Kluwer Academic, Dordrecht/Boston, 1999).
24. Liseau, R. & Artymowicz, P. High sensitivity search for molecular gas in the Beta Pic disk. On the low gas-to-dust mass ratio of the circumstellar disk around Beta Pictoris. *Astron. Astrophys.* **334**, 935–942 (1998).

Acknowledgements

We thank Gopal-Krishna, G. Pineau des Forêts and S. Cabrit for discussions, and R. Kurucz for providing a model of the β Pic stellar atmosphere. The work at Johns Hopkins University was supported by NASA; the work at the Institut d'Astrophysique de Paris and at LAM was supported by CNES. The work reported here is based on data obtained by the NASA-CNES-CSA FUSE mission, operated by the Johns Hopkins University.

Correspondence and requests for materials should be addressed to A.L.d.E. (e-mail: lecaveli@iap.fr).

Origin of the Moon in a giant impact near the end of the Earth's formation

Robin M. Canup* & Erik Asphaug†

* Department of Space Studies, Southwest Research Institute, 1050 Walnut Street, Suite 426, Boulder, Colorado 80302, USA

† Department of Earth Sciences, University of California, Santa Cruz, California 95064, USA

The Moon is generally believed to have formed from debris ejected by a large off-centre collision with the early Earth^{1,2}. The impact orientation and size are constrained by the angular momentum contained in both the Earth's spin and the Moon's orbit, a quantity that has been nearly conserved over the past 4.5 billion years. Simulations of potential moon-forming impacts now achieve resolutions sufficient to study the production of bound debris. However, identifying impacts capable of yielding the Earth–Moon system has proved difficult^{3–6}. Previous works^{4,5} found that forming the Moon with an appropriate impact angular momentum required the impact to occur when the Earth was only about half formed, a more restrictive and problematic model than that originally envisaged. Here we report a class of impacts that yield an iron-poor Moon, as well as the current masses and angular momentum of the Earth–Moon system. This class of impacts involves a smaller—and thus more likely—object than previously considered viable, and suggests that the Moon formed near the very end of Earth's accumulation.

The strength of the impact hypothesis over alternative models rests on its ability to account for (1) the initial ~5-hour terrestrial day implied by the Earth–Moon system angular momentum ($L_{E-M} = 3.5 \times 10^{41}$ g cm² s⁻¹); and (2) the ejection of sufficient iron-depleted material into orbit to yield the Moon, which has an unusually small metallic core comprising ≲3% of its mass⁷. Many works^{3–5,8–12} have modelled potential moon-forming impacts, most using a method known as smooth particle hydrodynamics, or SPH¹³. In SPH, objects are evolved in time by estimating their state and dynamical variables at discrete points that are smoothed over spherical overlapping kernel functions. This lagrangian technique requires no underlying grid, and is well suited to intensely deforming systems evolving within mostly empty space.

Recent SPH simulations^{4,5} identify two classes of impacts capable of placing sufficient mass into orbit to yield the Moon, but neither is entirely satisfactory. The first involves impacts with angular

momentum, L_{imp} , much greater than $L_{\text{E-M}}$, typically by a factor of 2. A significant dynamical event, for example another giant impact, would then be required to decrease the Earth–Moon system angular momentum by such a large amount. The second involves impacts with $L_{\text{imp}} \approx L_{\text{E-M}}$, but with a total mass (impactor plus target) of only $M_{\text{T}} \approx 0.65M_{\oplus}$, where $M_{\oplus}$ is the mass of the Earth (5.98×10^{27} g). In this ‘early Earth’ scenario, the Earth is only partially accreted when the Moon forms and must subsequently gain about $0.35M_{\oplus}$, with the later growth involving sufficiently small and numerous impacts that the system angular momentum is not substantially altered⁵. If, during this period, the Moon also accumulated even an approximately proportionate amount of

material, it would have gained excessive amounts of iron. The ratio of impact rate onto the Moon versus that with the Earth is $N_1 = f_{\text{M}}R_{\text{M}}^2/f_{\oplus}R_{\oplus}^2$, where $f_{\text{M}}/f_{\oplus}$ is the ratio of the gravitational focusing factor of the Moon to that of the Earth, and R_{M} and $R_{\oplus}$ are the lunar and terrestrial radii. Reasonable values of impactor velocity yield¹⁴ $0.03 < N_1 < 0.074$; assuming an impacting population with a terrestrial abundance of iron implies that the Earth could accrete only about $0.05M_{\oplus}$ before impacts with the Moon delivered an amount of iron equal to the upper limit on the mass of the lunar core.

It has recently been shown⁶ that simulations with a fixed ratio $\gamma = M_{\text{imp}}/M_{\text{T}}$ of the impactor-to-total mass yield a maximum

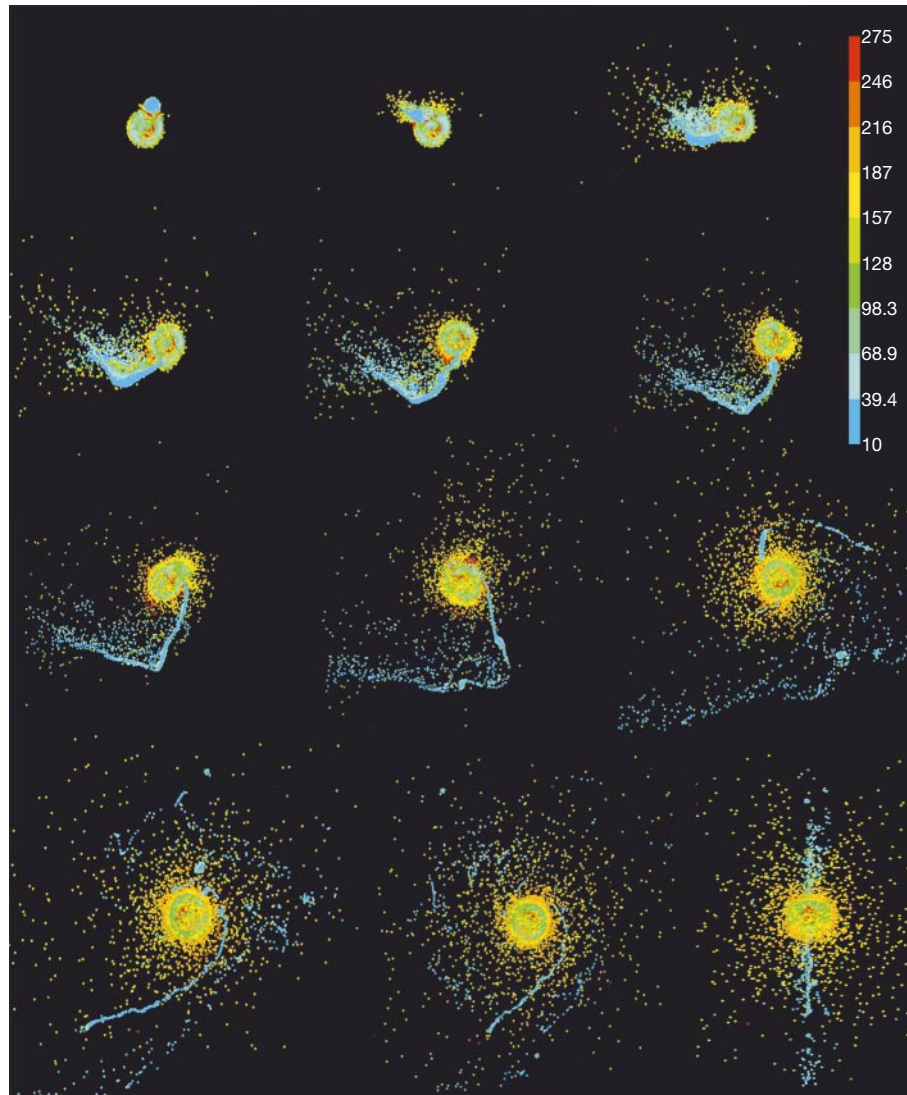


Figure 1 Time series of a Moon-forming impact simulation. Results are shown looking down onto the plane of the impact at times $t = 0.3, 0.7, 1.4, 1.9, 3, 3.9, 5, 7.1, 11.6, 17$ and 23 hours (from left to right); the last frame is $t = 23$ hours viewed on-edge. Colour scales with internal energy (shown on the colour bar in units of $6.67 \times 10^8 \text{ erg g}^{-1}$), so that blue and dark green represents condensed matter, and red particles signify either the expanded phase or a hot, high-pressure condensed phase; pressures at intermediate energies are computed by an interpolation between the Tillotson¹⁵ condensed and expanded phases. We form initial impactors and targets in hydrostatic equilibrium by pre-colliding smaller bodies together at zero incidence, resulting in realistically evolved internal energies, stratified densities (basalt mantle + iron core) and consistent pressures. Each particle’s internal energy is evolved due to the effects of expansion/compression and shock dissipation, with the latter represented by

artificial viscosity terms that are linear and quadratic in the velocity divergence of converging particles; effects of mechanical strength and radiative transfer are ignored. The momentum of each particle is evolved due to pressure, viscous dissipation and gravity. Gravity is computed using a binary tree algorithm, reducing the N^2 calculation of particle–particle attractions into an $N \log N$ calculation²⁵. We use a beta spine kernel to define the spatial distribution of material represented by each SPH particle. The scale of each particle, h , is automatically adjusted to cause overlap with a minimum of 40 other particles, ensuring a ‘smoothed’ distribution of material even in low-density regions. The code is explicit, requiring a Courant-limited timestep $\Delta t < (c/h)$ where c is the sound speed. For a full description of the technique, see ref. 26, from whose efforts our present algorithm derives.

orbiting mass when the impact parameter $b \approx 0.8$ ($b = L_{\text{imp}}/L_{\text{graz}} = 0$ and 1 for head-on and grazing collisions, and no collision occurs for $b > 1$). This result is independent of the total colliding mass M_T . Accordingly, the maximum yield for an $L_{\text{imp}} \approx L_{\text{E-M}}$ and $M_T \approx M_{\oplus}$ impact should be achieved when $b \approx L_{\text{E-M}}/L_{\text{graz}} \approx 0.8$, with the angular momentum of a grazing impact, L_{graz} , given by (assuming an impact velocity, v_{imp} , equal to the mutual escape velocity of the colliding bodies, v_{esc}):

$$L_{\text{graz}} = [3/(4\pi\rho)]^{1/6} \sqrt{2Gf(\gamma)M_T^{5/3}} \quad (1)$$

where G is the gravitational constant, ρ is the average target/impactor density, and $f(\gamma) \equiv \gamma(1-\gamma)[\gamma^{1/3} + (1-\gamma)^{1/3}]^{1/2}$. Setting $L_{\text{E-M}}/L_{\text{graz}} = 0.8$ and $M_T = M_{\oplus}$ yields a predicted optimal impactor-to-total mass ratio of $\gamma \approx 0.1$.

Colliding objects with $\gamma < 0.12$ had been ruled out as lunar-forming candidates in early low-resolution studies⁹, because they appeared to produce overly iron-rich disks. However, for those simulations with $N = 3,000$ particles, a single iron particle contained a mass comparable to the upper limit on the lunar core. We now use increased resolution to revisit the smaller-impactor scenario, with ejected debris described by 10^2 – 10^3 particles. With higher resolution, (1) the shock wave is better resolved, and it is the shock and its release that are responsible for the flow field and the expansion of vapour at the impactor/target boundary; (2) errors implicit in sparse particles are reduced, as SPH only has meaning as an interpolation technique over sets of overlapping particles; and (3) the granularity of the gravitational potential of orbiting debris is lessened. Such issues were recognized previously, but $\gamma < 0.12$ impacts have not been re-examined until now.

Figure 1 shows a time series of a simulation with $N = 30,000$ particles (run 24; see Table 1), in which the mass ratio $\gamma = 0.11$. At the end of this simulation, the planet contains $0.99M_{\oplus}$ and the mass of orbiting debris is $M_D = 1.68M_M$ (where $M_M = 0.012M_{\oplus}$ is a lunar mass). Of the orbiting mass, 2.0% is iron, with 80% of this iron located within the Earth's Roche limit (the distance interior to which a strengthless, self-gravitating object would be torn apart by Earth's tidal forces), where $a_{\text{Roche}} = 2.9R_{\oplus}$ for lunar and terrestrial densities. The orbiting material consists of a cooler component that forms a flattened equatorial disk, and a hot, pressure-supported component that comprises a vertically thick cloud enshrouding the planet.

Our simulations (like those of refs 8–10) rely upon the Tillotson¹⁵ equation of state (EOS) to calculate pressure. The Tillotson EOS is a relatively simple EOS developed for strong shocks in metals and geologic media. Direct comparisons¹⁶ between simulations using the code employed here and that of Cameron⁴ utilizing a more sophisticated EOS, known as ANEOS¹⁰, show general similarities in the predicted amount of mass and angular momentum placed into orbit, with the latter consistent to $\sim 10\%$. However, these comparisons showed significant differences in the implied physical state of the orbiting material. Unlike the Tillotson EOS, ANEOS handles mixed phases; however, in its current rendition ANEOS treats all vapour as monatomic species¹⁷. The entropy required for vaporization of molecular species such as mantle rock is therefore greatly overestimated by ANEOS, and a remedy to this is being undertaken by Melosh¹⁸. This EOS problem or a related error may contribute to the perplexing near-zero temperatures in the mantle and core of the final protoearth in Cameron's recent simulations⁴. For now, we

Table 1 Parameters and results of impact simulations

Run	$M_T/M_{\oplus}$	γ	$L_{\text{imp}}/L_{\text{graz}}$	$L_{\text{imp}}/L_{\text{E-M}}$	$L_D/L_{\text{E-M}}$	$L_{\text{final}}/L_{\text{E-M}}$	M_{esc}/M_M	M_{Fe}/M_D	M_D/M_M	$M_D/M_M > a_{\text{Roche}}$	M/M_M
1	1.018	0.08	0.98	1.07	0.27	0.80	1.11	0.32	1.50	0.86	1.13
3	1.018	0.09	0.88	1.07	0.29	1.00	0.48	0.05	1.50	0.82	1.30
9	1.018	0.09	0.90	1.10	0.25	0.90	0.94	0.12	1.37	0.82	1.04
17	1.018	0.09	0.94	1.15	0.30	0.92	0.98	0.16	1.51	1.04	1.40
18	1.018	0.09	0.98	1.20	0.30	0.88	1.28	0.24	1.62	1.02	1.29
54	0.980	0.097	0.84	1.04	0.30	0.97	0.48	0.01	1.32	0.92	1.32
53	0.980	0.097	0.90	1.12	0.24	0.93	0.89	0.07	1.31	0.74	1.02
2	1.019	0.10	0.80	1.07	0.22	1.00	0.53	0.01	1.18	0.66	0.96
14	1.109	0.10	0.82	1.10	0.28	1.03	0.48	0.01	1.40	0.78	1.32
23	1.019	0.10	0.84	1.12	0.30	1.05	0.50	<0.01	1.35	0.87	1.35
8	1.019	0.10	0.85	1.15	0.28	1.07	0.52	0.04	1.32	0.76	1.30
16	1.019	0.10	0.87	1.17	0.23	0.98	0.91	0.04	1.36	0.50	0.84
15	1.019	0.10	0.89	1.20	0.27	1.00	0.98	0.05	1.50	0.75	1.03
5	1.019	0.107	0.75	1.07	0.12	0.99	0.61	0.02	0.78	0.20	0.36
21	1.019	0.107	0.77	1.10	0.19	1.03	0.56	0.02	1.09	0.51	0.74
12	1.019	0.107	0.80	1.15	0.26	1.07	0.55	0.01	1.27	0.78	1.27
59	0.990	0.107	0.82	1.13	0.31	1.06	0.45	0.01	1.59	0.97	1.49
6	1.019	0.107	0.82	1.18	0.34	1.10	0.55	0.01	1.69	1.16	1.69
31	1.012	0.107	0.82	1.16	0.30	1.07	0.67	0.01	1.52	0.73	1.33
22	1.019	0.107	0.84	1.20	0.32	1.12	0.54	0.03	1.47	1.04	1.47
^c 60	0.990	0.107	0.90	1.23	0.33	1.04	0.88	0.09	1.82	1.00	1.38
11	1.019	0.107	0.90	1.29	0.31	1.08	1.00	0.05	1.71	0.96	1.29
4	1.019	0.11	0.73	1.07	0.09	0.99	0.59	<0.01	0.66	0.06	0.25
10	1.019	0.11	0.75	1.10	0.14	1.02	0.59	0.01	0.92	0.22	0.42
19	1.019	0.11	0.78	1.15	0.24	1.07	0.56	0.01	1.22	0.70	1.04
25	0.980	0.11	0.80	1.11	0.28	1.03	0.65	<0.01	1.31	0.89	1.31
^a 36	1.019	0.11	0.82	1.20	0.22	1.10	0.76	<0.01	1.23	0.57	1.00
^b 35	1.019	0.11	0.82	1.20	0.36	1.11	0.71	<0.01	1.68	1.26	1.68
20	1.019	0.11	0.82	1.20	0.33	1.12	0.51	<0.01	1.60	1.12	1.58
^c 24	1.019	0.11	0.82	1.20	0.34	1.12	0.50	0.02	1.68	0.97	1.48
^d 34	1.017	0.11	0.82	1.20	0.31	1.12	0.64	0.02	1.50	0.81	1.50
26	0.980	0.11	0.82	1.13	0.34	1.05	0.64	0.01	1.59	1.19	1.59
27	0.980	0.11	0.84	1.15	0.34	1.07	0.58	0.01	1.65	1.27	1.65
28	0.970	0.115	0.78	1.10	0.19	1.01	0.74	0.01	1.09	0.54	0.78
29	0.970	0.115	0.80	1.13	0.31	1.05	0.66	<0.01	1.50	1.01	1.50
30	0.970	0.115	0.82	1.16	0.36	1.08	0.63	0.01	1.75	1.19	1.75

At the end of each run, we classify particles as escaping (positive kinetic + potential energies), orbiting, or within the planet. A particle is considered to be orbiting if its total energy is negative, and the z-component of its angular momentum exceeds that of a circular orbit at the surface of the planet. We use an iterative analytic method^{6,16} to solve for the mass and radius of the final planet in order to maintain consistency across different simulations. First, an estimate of the fraction of mass contained in the planet is used to calculate the amount of escaping and bound debris; this yields an updated value for the planet's mass and radius, assuming a terrestrial bulk density. Results shown are at approximately 24 hours post-impact; for resolutions used here, numerically induced spreading in the disk²⁴ should be minimal on this timescale. M_D and L_D are the mass and angular momentum in orbiting debris; $M_D/M_M > a_{\text{Roche}}$ is the orbiting mass in lunar masses that has an equivalent orbit exterior to a_{Roche} . M_{Fe} is the mass of orbiting iron, and L_{final} is L_{imp} minus angular momentum of escaping debris. The nominal resolution is $N = 2 \times 10^4$ particles; runs with $N = 3 \times 10^4$, 1×10^4 , 3×10^4 and 6×10^4 particles are indicated with a, b, c and d superscripts, respectively. In all cases $v_{\text{imp}} = v_{\text{esc}}$, with no pre-impact spin. The final column shows the predicted mass of the moon in lunar masses (M/M_M) that would result (see text). Boldface indicates cases where $M/M_M \geq 1$ and $M_{\text{Fe}}/M_D \leq 0.03$.

consider the Tillotson EOS to be more robust. Although it cannot reliably predict melt fraction, it renders shock compression and release quite well and these, as well as gravitational torques, are of primary importance in mobilising the ejected matter into orbit.

The 36 simulations listed in Table 1 were parameterized so that the resulting planet-debris system had a mass within 5% of M_{E-M} ($M_{E-M} \equiv M_{\oplus} + M_M = 1.012M_{\oplus}$), and a final angular momentum within about 10% of L_{E-M} . In 4.5 billion years, a system decrease of less than $0.1L_{E-M}$ would occur⁶ owing to direct solar tides raised on the Earth if we assume a terrestrial tidal dissipation factor, Q , larger than $Q \approx 5$; currently, $Q \approx 12$.

Typically, the majority of orbiting debris has an equivalent circular orbit outside the Roche limit; none of our simulations produced massive disks located entirely within a_{Roche} . Most of the orbiting material originates from the impacting object rather than the target (the Earth), as seen earlier^{3-4,8-11}. For the boldface cases in Table 1, the fraction of orbiting mass originating from the impactor, M_i/M_D , ranges from 0.6 to 0.74; for material with angular momentum sufficient to orbit exterior to a_{Roche} , this fraction is higher, with $M_i(>a_{\text{Roche}})/M_D(>a_{\text{Roche}}) = 0.72$ to 0.88. This implies that the Moon will accrete from primarily impactor-originating material, with a possible target contribution of up to tens of per cent by mass. However the relative amount of ejecta from the impactor versus target may be affected by the EOS treatment of vaporization¹⁶.

Larger impactors ($\gamma > 0.115$) can produce massive disks, but yield an Earth–Moon system with too much angular momentum. For a fixed L_{imp} and M_T , a larger γ requires a smaller impact parameter b , and under this constraint a larger γ results in less orbiting debris. Recent simulations⁵ use $\gamma = 0.3$, and for this value $M_D \leq 0.05M_M$ for $L_{\text{imp}} \approx L_{E-M}$ and $M_T \approx M_{E-M}$. Unless a progressively smaller total colliding mass is considered (for example, for $\gamma = 0.3$, M_T must be reduced to about $0.65M_{\oplus}$; refs 4–6), it appears impossible to produce an appropriate protolunar disk for larger γ values and $L_{\text{imp}} \approx L_{E-M}$. On the other hand, smaller impactors ($0.08 \leq \gamma \leq 0.09$) produce disks that are too iron-rich, because their impact parameter b must be near grazing and so too much of the impactor core remains in orbit. We note however that in nearly all cases, the majority of orbiting iron is located interior to a_{Roche} , which could lead to a less iron-rich Moon than is implied by the bulk composition of the orbiting material. Still lower γ -values cannot deliver $L_{\text{imp}} \geq L_{E-M}$ for $v_{\text{imp}} \approx v_{\text{esc}}$, unless the planet is already spinning rapidly in the same sense as the collision.

Four simulations of the impact shown in Fig. 1 with varied resolutions predict a mean orbiting mass of $\langle M_D \rangle = 1.62M_M \pm 0.15$ and a mean angular momentum in orbiting material of $\langle L_D \rangle = 0.335L_{E-M} \pm 0.036$. In some of the lower-resolution runs (for example, runs 35 and 20) we find that a large intact moon forms as a direct result of the impact (also observed earlier^{9,10}), whereas

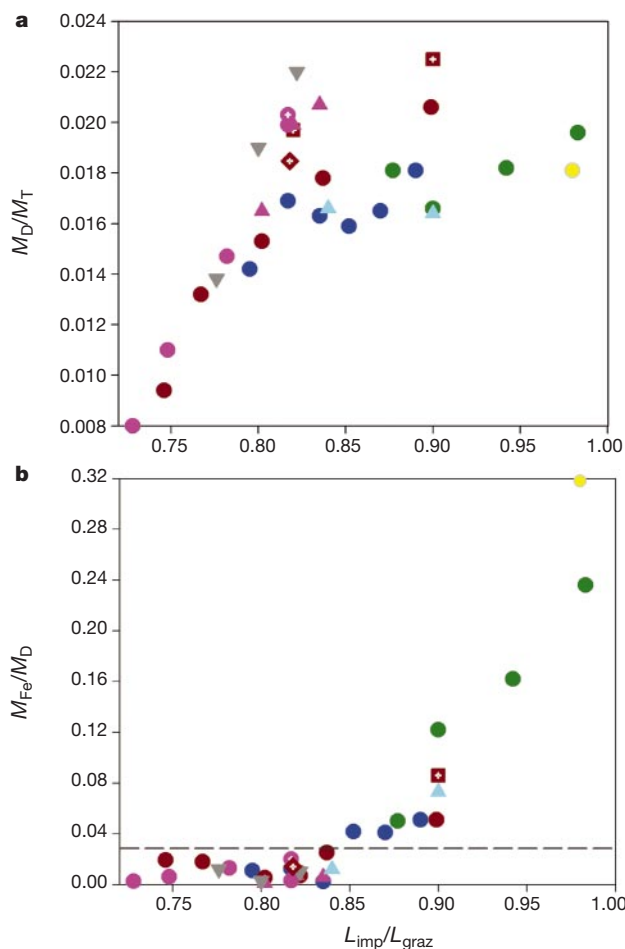


Figure 2 Scaled smooth particle hydrodynamics simulation results. **a**, The orbiting mass, M_D , scaled by the total colliding mass, M_T , as a function of $b = L_{\text{imp}}/L_{\text{graz}}$, where b is the impact parameter, with $b = 1$ for a grazing impact. **b**, The mass fraction of iron in the orbiting mass, M_{Fe}/M_D versus $L_{\text{imp}}/L_{\text{graz}}$. A disk iron composition equal to the approximate mass fraction of the lunar core ($M_{\text{Fe}}/M_D = 0.03$) (ref. 7) is shown by the horizontal dashed line. Colours vary with the impactor-to-total-mass ratio, γ , with yellow, green,

cyan, blue, dark red, magenta and grey corresponding respectively to $\gamma = 0.08, 0.09, 0.097, 0.10, 0.107, 0.11$ and 0.115 . Symbols vary with total mass: $M_T = 0.97M_{\oplus}$ (inverted-triangle), $M_T = 0.98M_{\oplus}$ (triangle), $M_T = 0.99M_{\oplus}$ (square), $M_T = 1.01M_{\oplus}$ (diamond) and $M_T = 1.02M_{\oplus}$ (circle). Simulations with $N = 30,000$ have a white cross-hatch within their respective symbols; all others have $N = 20,000$.

higher-resolution runs of the same impact (runs 24 and 34) produce a debris disk. The number of particles dictates the coarseness of the gravity potential, especially in a drawn-out bar of material. When the scale of particles is larger than the wavelength of the physical instability in the continuum those particles represent, collapse might ensue, whereas a more finely resolved system would be drawn apart by tidal shearing. This artefact of low-resolution was noticed in models¹⁹ for the tidal disruption of comet Shoemaker–Levy 9.

Figure 2 shows scaled data from Table 1. General trends with increasing impact parameter are seen across runs with different total masses and γ values; there is also a dependence of the maximum yield on γ . For $b > 0.85$, orbiting material contains more than 3% iron by mass, with this fraction increasing progressively until a disk with a similar iron-mantle composition to that of the impactor is achieved for grazing impacts⁶.

Models of protolunar disk accretion^{20–22} find that a large moon forms at a characteristic distance of about $1.2a_{\text{Roche}}$, with a mass that is a function of the initial disk mass and angular momentum:

$$M \approx 1.9L_D/\sqrt{GM_\oplus a_{\text{Roche}}} - 1.1M_D - 1.9M_{\text{esc}} \quad (2)$$

where M_{esc} is the amount of escaping material during accretion. Using $M_{\text{esc}} = 0.05M_D$ (refs 21, 22), we estimate the moon mass that would result in the final column of Table 1; simulations for which this value is at least a lunar mass with $M_{\text{Fe}}/M_D \leq 0.03$ are shown in boldface.

The successful impacts involve an impactor-to-target mass ratio $\gamma \approx 0.1$ – 0.11 , or an impactor with a mass of ~ 6 – 6.5×10^{26} g—the mass of Mars. This class of impacts represents the least restrictive impact scenario, requiring little or no dynamical modification of the Earth–Moon system after the moon-forming impact. Problems associated with a period of extended terrestrial growth subsequent to the event are avoided; in addition, the smaller impact we advocate is more likely than the 2–3 times more massive impactors postulated by recent works^{4,5}; this is because in collisional populations small objects are more common than large ones (the number of objects, dN , in a mass range dm is typically proportional to $dN \propto m^{-q}dm$, with²³ $q \approx 1.5$ to 1.8). It is, in retrospect, interesting that the Mars-mass impactor that now appears to be the most promising Moon-forming candidate is essentially that originally proposed¹, a decade before the first Moon-forming impact simulations. □

Received 23 April; accepted 12 June 2001.

1. Cameron, A. G. W. & Ward, W. R. The origin of the Moon. *Lunar Sci.* **7**, 120–122 (1976).
2. Hartmann, W. K. & Davis, D. R. Satellite-sized planetesimals and lunar origin. *Icarus* **24**, 504–515 (1975).
3. Cameron, A. G. W. The origin of the Moon and the single impact hypothesis V. *Icarus* **126**, 126–137 (1997).
4. Cameron, A. G. W. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 133–144 (Univ. Arizona Press, Tucson, 2000).
5. Cameron, A. G. W. From interstellar gas to the Earth–Moon system. *Meteor. Planet. Sci.* **36**, 9–22 (2001).
6. Canup, R. M., Ward, W. R. & Cameron, A. G. W. A scaling law for satellite-forming impacts. *Icarus* **150**, 288–296 (2001).
7. Hood, L. L. & Zuber, M. T. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 397–409 (Univ. Arizona Press, Tucson, 2000).
8. Benz, W., Slattery, W. L. & Cameron, A. G. W. The origin of the Moon and the single impact hypothesis I. *Icarus* **66**, 515–535 (1986).
9. Benz, W., Slattery, W. L. & Cameron, A. G. W. The origin of the Moon and the single impact hypothesis II. *Icarus* **71**, 30–45 (1987).
10. Benz, W., Cameron, A. G. W. & Melosh, H. J. The origin of the Moon and the single impact hypothesis III. *Icarus* **81**, 113–131 (1989).
11. Cameron, A. G. W. & Benz, W. The origin of the Moon and the single impact hypothesis IV. *Icarus* **92**, 204–216 (1991).
12. Melosh, H. J. & Kipp, M. E. Giant impact theory of the Moon's origin: first 3-D hydrocode results. *Lunar Sci.* **20**, 685–686 (1989).
13. Lucy, L. B. A numerical approach to the testing of the fission hypothesis. *Astron. J.* **82**, 1013–1024 (1977).
14. Stewart, G. R. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 217–223 (Univ. Arizona Press, Tucson, 2000).
15. Tillotson, J. H. Metallic equations of state for hypervelocity impact. Report No. GA-3216, July 18 (General Atomic, San Diego, California, 1962).

16. Canup, R. M. & Asphaug, E. Outcomes of planet-scale collisions. *Lunar Sci.* [CD-ROM] **32**, (2001).
17. Melosh, H. J. & Pierazzo, E. Impact vapor plume expansion with realistic geometry and equation of state. *Lunar Sci.* **28**, 935 (1997).
18. Melosh, H. J. A new and improved equation of state for impact computations. *Lunar Sci.* **31**, 1903 (2000).
19. Asphaug, E. & Benz, W. Size, density, and structure of comet Shoemaker–Levy 9 inferred from the physics of tidal breakup. *Icarus* **121**, 225–248 (1996).
20. Ida, S., Canup, R. M. & Stewart, G. Formation of the Moon from an impact-generated disk. *Nature* **389**, 353–357 (1997).
21. Kokubo, E., Canup, R. M. & Ida, S. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 145–163 (Univ. Arizona Press, Tucson, 2000).
22. Kokubo, E., Makino, J. & Ida, S. Evolution of a circumterrestrial disk and formation of a single Moon. *Icarus* **148**, 419–436 (2001).
23. Greenberg, R. in *Origin and Evolution of Planetary and Satellite Atmospheres* (eds Atreya, S. K., Pollack, J. B. & Matthews, M. S.) 137–164 (Univ. Arizona Press, Tucson, 1989).
24. Nelson, A., Benz, W., Adams, F. & Arnett, D. Dynamics of circumstellar disks. *Astrophys. J.* **502**, 342–371 (1998).
25. Hernquist, L. & Katz, N. TREESPH—A unification of SPH with the hierarchical tree method. *Astrophys. J. Suppl.* **70**, 419–446 (1989).
26. Benz, W. in *Proc. NATO Adv. Res. Workshop on Numerical Modelling of Nonlinear Stellar Pulsations* (ed. Buchler, J. R.) 1–54 (Kluwer Academic, Boston, 1990).

Acknowledgements

We wish to thank Southwest Research Institute's Internal Research program for its support of development efforts for the methods utilized here, D. Terrell for securing a portion of the computational time, and P. Tamblin for aid with some of the analysis software. A review by A. Halliday and comments provided by W. Ward, C. Agnor, D. Korycansky and R. Mihar helped to improve the paper. This research was supported by the National Science Foundation and NASA.

Correspondence and requests for materials should be addressed to R.M.C. (e-mail: robin@boulder.swri.edu).

Sub-Planck structure in phase space and its relevance for quantum decoherence

Wojciech Hubert Zurek

Theory Division, T-6, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Heisenberg's principle¹ states that the product of uncertainties of position and momentum should be no less than the limit set by Planck's constant, $\hbar/2$. This is usually taken to imply that phase space structures associated with sub-Planck scales ($\ll \hbar$) do not exist, or at least that they do not matter. Here I show that this common assumption is false: non-local quantum superpositions (or 'Schrödinger's cat' states) that are confined to a phase space volume characterized by the classical action A , much larger than $\hbar$, develop spotty structure on the sub-Planck scale, $a = \hbar^2/A$. Structure saturates on this scale particularly quickly in quantum versions of classically chaotic systems—such as gases that are modelled by chaotic scattering of molecules—because their exponential sensitivity to perturbations² causes them to be driven into non-local 'cat' states. Most importantly, these sub-Planck scales are physically significant: a determines the sensitivity of a quantum system or environment to perturbations. Therefore, this scale controls the effectiveness of decoherence and the selection of preferred pointer states by the environment^{3–8}. It will also be relevant in setting limits on the sensitivity of quantum meters.

One of the characteristic features of classical chaos is the evolution of the small-scale structure in phase-space probability distributions. As a consequence of the exponential sensitivity to initial conditions, an initially regular 'patch' in phase space with a

characteristic size Δ will, after a time t , develop structure on the scale

$$z \approx \Delta \exp(-\Lambda t) \quad (1)$$

where Λ is, in effect, the Lyapunov exponent^{9,10}. This is a consequence of the exponential stretching and of the conservation of phase-space volume in reversible hamiltonian evolutions.

What happens with small-scale structure in quantum versions of classically chaotic systems? We shall investigate this question using the Wigner function $W(x, p)$ —the closest quantum analogue of the classical phase-space distribution¹¹:

$$W(x, p) = \frac{1}{2\pi\hbar} \int \exp(ip y/\hbar) \rho\left(x - \frac{y}{2}, x + \frac{y}{2}\right) dy \quad (2)$$

Above, ρ is the density operator, x is position and p is momentum. When the quantum state is a pure $|\psi\rangle$, the integrand becomes $\exp(ip y/\hbar) \psi^*(x + y/2) \psi(x - y/2)$.

In a quantum system the smallest spatial structures in the wavefunction $\psi(x)$ will be set by the highest phase-space frequencies available. Given a finite energy, $\psi(x)$ and, hence, $W(x, p)$ cannot reach scales as small as the exponential squeezing of equation (1) would eventually imply. Thus, in a quantum system, there must be a scale below which structure should not appear. That some limit must exist was apparent some time ago: Berry and Balazs (who, as we shall see below, have an enviable record of correctly anticipating aspects of quantum chaos) have conjectured^{12,13} that structure

saturates on scales given by the Planck constant. If this were the case, W should be smooth on scales small compared with $2\pi\hbar$.

I shall show that copious structure appears in the Wigner function W on much smaller sub-Planck scales associated with the action of the order of

$$a \approx \hbar \times \frac{\hbar}{A} \quad (3)$$

and explain its origin. Most importantly, I shall demonstrate that a has physical consequences: it determines the sensitivity of the system (or of the environment) to decoherence. Above, A is the classical action of the system, given approximately by the product of the range of effective support of its state (the phase space envelope within which Wigner function is significantly different from zero) in position $\Delta x = L$ and momentum $\Delta p = P$:

$$A \approx L \times P \quad (4)$$

The values of L and P are in turn set by the available energy E , and by the form of the potential $V(x)$. That is, $P \leq \sqrt{2mE}$, $E - V(L) \geq 0$, and so on, which determine the effective support of the probability distribution in phase space. We shall eventually see that many of the calculations can be carried out using the state vector of the system in the appropriate representation. Nevertheless, intuitive understanding of the significance of the sub-Planck scale a is easiest to attain starting with a more comprehensive view of the phase-space image of quantum states afforded by the Wigner representation.

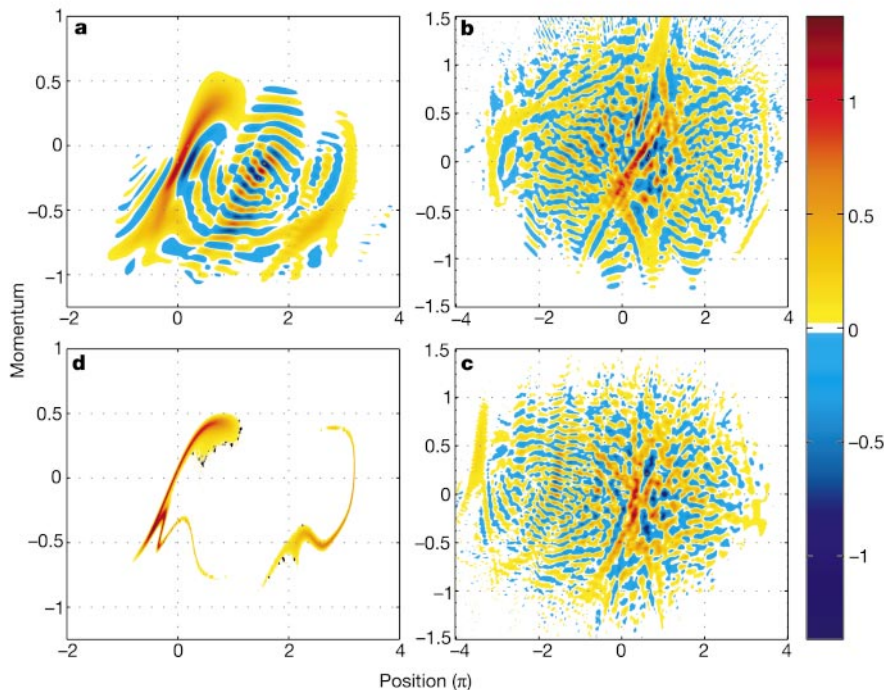


Figure 1 Snapshots of the quantum Wigner distribution and of the classical probability density in phase space of an evolving chaotic system. This system has the hamiltonian $H = (p^2/2m) - \kappa \cos(x - l \sin t) + ax^2/2$. For parameters $m = 1$, $\kappa = 0.36$, $l = 3$ and $a = 0.01$ the system exhibits chaos²⁴ with Lyapunov exponent $\Lambda \approx 0.2$. Initial probability density was given by the same gaussian in both the quantum cases (**a–c**; $\hbar = 0.16$) and the classical case (**d**). Structure on sub-Planck scales appears and saturates quickly. Note the contrast between the ever decreasing smallest dimensions of the probability density in the classical case (**d**) and of the Wigner distribution at the corresponding time (**a**). Exponential shrinking of the smallest scales makes it impossible to simulate accurately (for example, reversibly) classical evolution much beyond the time shown in **d**. By contrast, structures in quantum Wigner distribution saturate on scales

$a \approx \hbar^2/A$ soon after $t \approx 20$ (which is of the order of the estimated t_h , equation (13a)). The scale of structure saturation of a state can be inferred from the volume of the domain containing it: smallest structures have dimensions $\delta_x = \hbar/P$ ($\delta_p = \hbar/L$), where P and L define the extent of the envelope of the effective support of the state ($P^2 \approx \langle p^2 \rangle - \langle p \rangle^2$, $L^2 \approx \langle x^2 \rangle - \langle x \rangle^2$). Action associated with the smallest structures will then be $a \approx \hbar^2/LP \approx \hbar^2/A$ in one spatial (two phase-space) dimensions. Generalization to the case of many dimensions is straightforward. One way to understand structure saturation is through the menu of wavevectors available in the system that have momenta restricted to $P \approx \sqrt{E}$, where E is the total energy. The complementary argument ($V(L) \approx E$, where $V(x)$ is the potential) can be made for the smallest scale of ‘corrugation’ of the Wigner distribution in p .

It is evident from Fig. 1 that, at least in chaotic systems, the structure on sub-Planck scales appears and saturates quickly. Let us now illustrate its origin. The smallest scales in phase space arise from interference. Consider, for instance, a familiar ‘Schrödinger cat’ coherent superposition of two minimum-uncertainty gaussians⁶:

$$W(x, p) = \frac{G(x + x_0, p) + G(x - x_0, p)}{2} + (\pi\hbar)^{-1} \exp\left(-\frac{p^2 \xi^2}{\hbar^2} - \frac{x^2}{\xi^2}\right) \cos\left(p \frac{2x_0}{\hbar}\right) \quad (5)$$

where Wigner functions of the two gaussians ‘east’ and ‘west’ of the centre are

$$G(x \pm x_0, p - p_0) = (\pi\hbar)^{-1} \exp\left(-\frac{(x \pm x_0)^2}{\xi^2} - \frac{(p - p_0)^2 \xi^2}{\hbar^2}\right) \quad (6)$$

The cosine term W_{WE} in equation (5) is a symptom of interference. Its ripples have a frequency proportional to the separation $L = 2x_0$ between the two peaks. When we define the frequency of the ripple pattern in momentum f_p through $\cos(Lp/\hbar) = \cos(f_p p)$, then $f_p = L/\hbar$. The ridges and valleys of such an interference pattern are always parallel to the line of sight between the two gaussian peaks. Thus, standing on top of one gaussian peak, one could still see the other peak through the valleys (and between the ridges) of the interference term, even though its envelope is a factor of two higher than either gaussian.

Coherent states form an overcomplete set. We can therefore express an arbitrary pure state as a superposition of coherent states. The smallest interference structures in such an expansion arise from pairs of coherent states separated by the whole range available to the system in phase space. Therefore, we expect smallest scales with

$$f_x = P/\hbar, \quad \text{or} \quad \delta_x = \hbar/P \quad (7)$$

$$f_p = L/\hbar, \quad \text{or} \quad \delta_p = \hbar/L \quad (8)$$

As an example, consider the compass state, a Schrödinger-cat-like superposition of four minimum-uncertainty gaussians, one pair located north and south of the common centre, the other east and west (see Fig. 2). The Wigner distribution is quadratic in the wavefunction. Therefore, W of any superposition can be reconstructed from the contributions corresponding to, at most, pairs of states. The structure of the Wigner distribution of a superposition of a pair of gaussians (equation (5)); note that this is reflected in the patterns of the sides of the square in Fig. 2) can be then used to infer W of the compass state:

$$W_{NWSE} = G_N + G_W + G_S + G_E)/4 \quad (9a)$$

$$+ (W_{NW} + W_{WS} + W_{SE} + W_{EN})/2 \quad (9b)$$

$$+ (W_{NS} + W_{EW})/2 \quad (9c)$$

in an obvious ‘geographical’ notation. The last line of equation (9),

$$W_{NS} + W_{EW} = (\pi\hbar)^{-1} \exp\left(\frac{p^2 \xi^2}{\hbar^2} - \frac{x^2}{\xi^2}\right) \left(\cos \frac{pL}{\hbar} + \cos \frac{pP}{\hbar}\right) \quad (10)$$

is of greatest interest. Above, I have assumed that the shapes of all the gaussians are identical, so that the exponential envelope in equation (9c) is common to both terms.

The resulting interference term (the centre of Fig. 2) has a chessboard pattern. The size of the single ‘tile’ can be obtained from zeros of the oscillatory factor of equation (10):

$$\cos(pL/\hbar) + \cos(pP/\hbar) = 2 \cos \frac{Px + Lp}{2\hbar} \cos \frac{Px - Lp}{2\hbar} \quad (11)$$

Individual squares, four per tile, are defined by zeros that occur

when $x = \pm \pi\hbar/2L$, $p = \pm \pi\hbar/2P$. The fundamental periodic tile has an area of

$$a = \frac{2\pi\hbar}{L} \times \frac{2\pi\hbar}{P} = (2\pi\hbar)^2/A \quad (12)$$

which, with the identification of $A = LP$, yields action associated with the smallest scales present in quantum phase space.

The above calculation shows by construction that a quantum state spread over a phase space of volume $A = LP$ can accommodate Wigner distribution structures as small as a . Such states arise naturally. Evolution will force almost any system (with a notable exception of a harmonic oscillator) into a ‘Schrödinger cat’ state—a coherent non-local superposition that, after the time it takes the wavefunction to spread over the phase-space volume A , inevitably develops an interference pattern on the scale given by equations (7–12).

It is not necessary to invoke chaos: after sufficient time, even an integrable nonlinear system may spread coherently throughout the available phase space, and, consequently, saturate small-scale

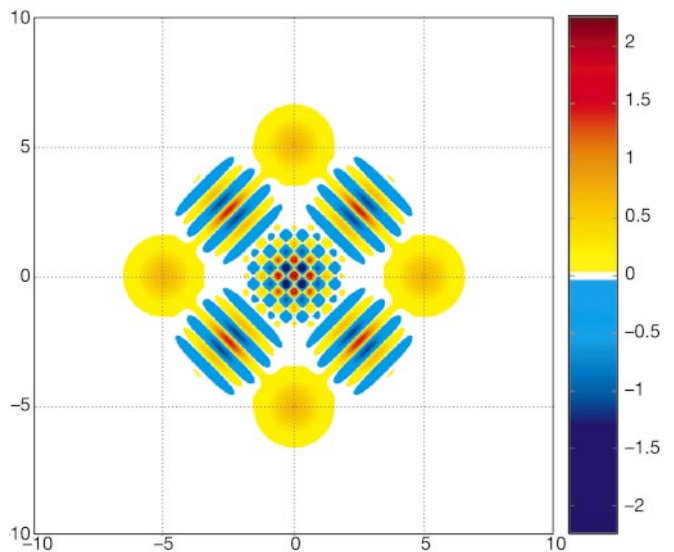


Figure 2 The compass state, equation (9). This is a superposition of minimum-uncertainty states ($|N\rangle, |W\rangle, |S\rangle, |E\rangle$) placed ‘north’, ‘...’, ‘east’, of the common centre. Units are selected so that the gaussians used are round and $\hbar = 1$. The form of the central interference pattern in W_{NWSE} can be inferred from the familiar structure of the superposition of two gaussians, equation (5), also apparent on the sides of the square above: the interference terms corresponding the north–south and east–west pairs superpose, creating a chessboard pattern. The area of individual tiles is set by the dimensions of the NWSE cross and corresponds to the classical action, A , of phase-space area of the effective support of its envelope, equation (12). The appearance of this interference pattern explains the origin of the structure saturation seen in Figs 1 and 3: a system that can be effectively confined to phase-space volume A cannot develop structure on scales smaller than a . Sensitivity of the compass state to perturbations is controlled by a . This is readily seen in the ‘sparse limit’, that is, when $L \gg \xi$, $P \gg \hbar/\xi$. For simplicity, consider a shift $\delta = \delta_x \hat{x} + \delta_p \hat{p}$ small compared with the sizes of the gaussians, $\delta_x \ll \xi$, $\delta_p \ll \hbar/\xi$. The square of the overlap of the original and displaced states $|\langle \psi_{NWSE} | \psi_{NWSE}^\delta \rangle|^2 = 2\pi\hbar \int W_{NWSE}(x, p) W_{NWSE}(x + \delta_x, p + \delta_p) dx dp$ is approximately $\langle N | N^2 \rangle + \dots + \langle E | E^2 \rangle \approx (\cos \delta_x P/2\hbar + \cos \delta_p L/2\hbar)^2/4$. To get this simple result we have ignored both the additive corrections (such as $\langle N | E^2 \rangle$) that are small in the sparse limit, and a multiplicative correction $\sim |\langle E | E^2 \rangle|^2$ that is very close to unity in the limit of small displacements $|\delta| \ll \hbar$. The striking kinship of this form of the overlap with the interference term, equations (10) and (11), is no accident: the magnitudes of the shifts that produce orthogonality are $\delta_x = 2\pi\hbar/P$, $\delta_p = 2\pi\hbar/L$. Thus, displacing the states by the size of the square in the central interference pattern defined by equations (11) and (12) suffices to cause decoherence, if the environment starts in the NWSE state.

structure. It is just that when the evolution is chaotic, such small scales will be attained faster, on a timescale given by¹⁴:

$$t_h = \Lambda^{-1} \ln \Delta p \chi / \hbar \quad (13a)$$

Above, Δp is the characteristic spread of the initial smooth probability distribution. χ characterizes the scale on which potential is significantly nonlinear. It is typically given by $\chi \approx \sqrt{(V''/V''')}$. Similar estimates are obtained from the formula

$$t_r = \Lambda^{-1} \ln A / \hbar \quad (13b)$$

deduced some time ago^{12,15}.

There is one more suggestive way to express saturation scale a : the number of distinct (orthogonal) states that can fit within phase space of volume A is $N = A/(2\pi\hbar)$. The structure we are discussing appears therefore on the scale $a \approx \hbar/N$. Here N is, in effect, the dimension of the available Hilbert space.

In accord with Heisenberg's principle, a quantum system cannot be localized to a sub-Planck volume in phase space. Hence, one might be tempted to dismiss sub-Planck scales as unphysical even if they appear in the Wigner distribution. In particular, if a state cannot be confined, by measurements, to a volume less than $\sim \hbar$, then one might expect that it will not be noticeably perturbed by displacements much smaller than $\sqrt{\hbar}$. I now show that this expectation is false, and that $a \ll \hbar$ plays a decisive role in determining the sensitivity of quantum systems (or of quantum environments) to perturbations: phase space displacements $\delta \approx \sqrt{a} \ll \sqrt{\hbar}$ shift a state with a dominant structure on scale a enough to make it orthogonal to—that is, distinguishable from—the unshifted original. Sensitivity to perturbations in turn sets the limit on the efficiency of decoherence.

There are two complementary aspects to this connection between a and decoherence. When a system with a sub-Planck scale in W is coupled to the environment, decoherence can be thought of as monitoring, by the environment, of some of its observables^{3–8}. Its effect—suppression of quantum coherence—can be traced to Heisenberg's principle: the observable that is complementary to the one monitored by the environment becomes less determined, in effect smearing Wigner distribution along the corresponding

phase-space direction⁸. When this smearing obliterates interference structures on scale a , coherence on the large scales corresponding to $A \approx \hbar^2/a$ will be suppressed, and the Schrödinger cat state will have lost its quantum nonlocality^{2,14,16}.

We turn now to a complementary aspect of the same story. It involves the situation when the state of the environment—the cause of decoherence—is a 'Schrödinger cat' spread over the phase-space region A . The environment entangles with the system, acting as a 'monitoring' apparatus^{3–8}. The sensitivity of the environment to perturbations is therefore of the essence. We shall test the sensitivity of such a Schrödinger cat environment by allowing it to interact with a Schrödinger cat system, assumed to be initially in a superposition of two perfect pointer states $\{|+\rangle, |-\rangle\}$. Pointer states, taken one at a time, perturb the state of the environment but do not entangle with it. However, each pointer state perturbs the environment differently³. Therefore, a system prepared in a general superposition state will leave a pointer-state-dependent imprint on the environment, and hence entangle with it. We shall show that the displacement $\delta \approx \sqrt{a}$ sets the size of the smallest perturbations distinguished by the environment, which in turn controls its ability to decohere the system. Thus, when $|+\rangle$ and $|-\rangle$ shift the state of the environment differently, the off-diagonal terms of the density matrix of the system will be suppressed by a factor $|\langle\epsilon_+|\epsilon_-\rangle|$, where

$$|\langle\epsilon_+|\epsilon_-\rangle|^2 = 2\pi\hbar \int W_+ W_- dx dp \quad (14)$$

Above, kets $|\epsilon_{\pm}\rangle$ and their Wigner distributions ($W_{\pm}$) represent states that evolve from the original state of the environment $|\epsilon\rangle$ through the interaction (induced, say, by a system–environment interaction hamiltonian such as $\sim(|+\rangle\langle+| - |-\rangle\langle-|)(\hbar/i)\partial_x$ with the states $|\pm\rangle$ of the system, respectively.

Let us now demonstrate how the behaviour of the magnitude of the overlap

$$|\langle\epsilon_+|\epsilon_-\rangle| = \left| \int \epsilon^*(x) e^{i\partial_p x/\hbar} \epsilon(x + \delta_x) dx \right| \quad (15)$$

is controlled by a . Above $\delta = (\delta_x, \delta_p)$ is the net displacement

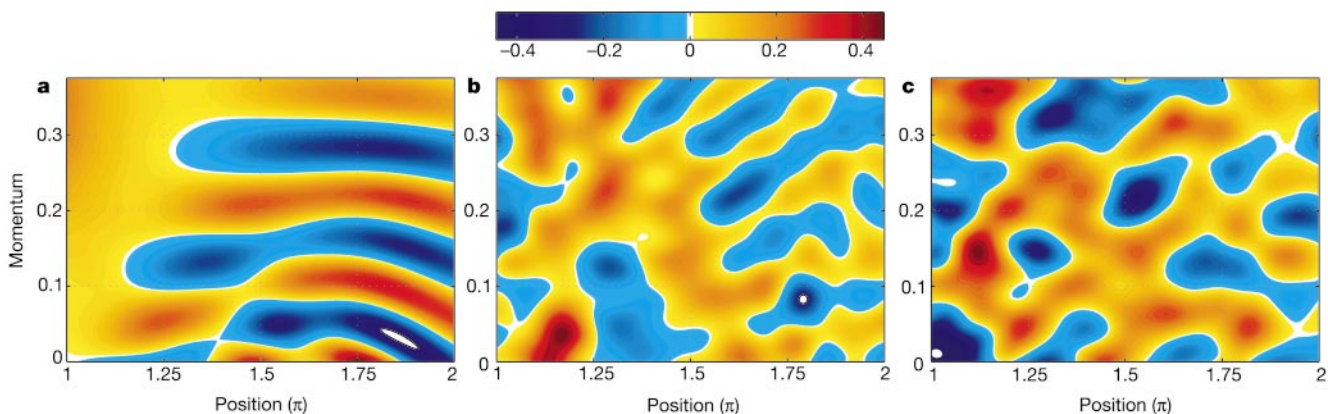


Figure 3 Snapshots of area $2\pi\hbar$ extracted from Fig. 1a–c. The smallest scales saturate on sub-Planck scale a , so that—as can be estimated from Fig. 1—the individual patches appear on the scale consistent with Figs 1 and 2 and equation (3). Such small phase-space substructure is physically significant. Displacement of the order of $\sqrt{a}$ (the size of a typical patch) suffices to make the perturbed state roughly orthogonal to its old (unperturbed) self. Thus, a sets the limit on the sensitivity of the state to perturbations, and is therefore relevant for decoherence. Note that the overlap of the original and displaced states will behave differently when the interference pattern is irregular (as it is here) rather than essentially periodic (as was the case in the compass state of Fig. 2). This can be understood by expressing the state as a superposition $|\psi\rangle = \sum_k \alpha_k |k\rangle$ of identical minimum-uncertainty gaussians $|k\rangle = G(x - x_k, p - p_k)$, each centred on x_k, p_k (see

equation (6)). In the sparse limit (that is, when the gaussians in the superposition of $|\psi\rangle$ do not overlap significantly) the overlap of the original and displaced states is approximately $|\langle\psi|\psi^\delta\rangle| \approx |\sum_k \alpha_k|^2 e^{i(\delta p x_k + \delta x p_k)/\hbar}| = |\sum_k w_k e^{i\phi_k}|$, where $w_k = |\alpha_k|^2$, $\sum_k w_k = 1$ are the weights (probabilities) of finding the system in different states $|k\rangle$. It is obvious that, when many sparsely distributed gaussians participate in $|\psi\rangle$, so that $w_k \approx 1/n$, the size of the overlap is given by the distance covered by an eventually random walk in a complex plane where individual steps have magnitude w_k and directions determined by phases ϕ_k . Hence, as ϕ_k become random, the overlap will rapidly decrease from unity to about $1/\sqrt{n}$, and will probably remain small. (Such sums are a standard way of recovering a gaussian probability distribution, and have already been studied in the context of decoherence some time ago⁴.)

between $|\epsilon_+\rangle$ and $|\epsilon_-\rangle$ corresponding to $W_+ = W(x + \delta_x^+, p + \delta_p^+)$ and $W_- = W(x + \delta_x^-, p + \delta_p^-)$, respectively. The displacement δ is the difference between the shifts caused by $|+\rangle$ and $|-\rangle$, $\delta = \delta^+ - \delta^- \neq 0$. The heuristic argument for the size of displacement that causes orthogonality $\langle \epsilon_+ | \epsilon_- \rangle \approx 0$, and hence decoherence, is easiest to follow when phrased in terms of Wigner functions. Suppose W_+ and W_- in equation (14) have a small-scale structure with patches of alternating sign, as seen in Figs 1–3. The integral of their product can reach a maximum value of unity only when W_+ and W_- are not shifted with respect to each other. For shifts small compared with the typical size of the patches the integrand will still be positive almost everywhere, but as the magnitude of the shift increases, the right-hand side of equation (14) will become small compared with unity. If the interference pattern in the Wigner distribution is periodic (as is the case in Fig. 2) the oscillation will be also periodic with a period related to the size of the fundamental ‘tile’. When, however, patches are random (as for the typical case illustrated in Fig. 3), the overlap—having decayed after a displacement $|\delta| \approx a$ —will not significantly recur. A more formal version of this heuristic argument is put forward and backed up by numerical simulations elsewhere (Z. Karkuszewski *et al.*, manuscript in preparation). A simple back-of-the-envelope calculation valid for a generalization of compass states (that is, when the state of the environment can be approximated by a ‘sparse’ collection of identically shaped minimum-uncertainty gaussians) is given in Fig. 3.

This intuitive picture, already supported by the example of Fig. 2, can be further confirmed by a general yet straightforward calculation based on equation (15). To simplify notation we consider the case when $\delta = (0, \delta_p)$, that is, when the net shift is aligned with one of the axes. (This assumption can be made with no loss of generality, as the axes in phase space can be rotated to align one of them with an arbitrary δ .) In that case:

$$\langle \epsilon_+ | \epsilon_- \rangle = \int |\epsilon(x)|^2 e^{i\delta_p x / \hbar} dx \quad (16)$$

This is a simple, general, and compelling result: suppression of the off-diagonal terms in the density matrix of the system is given by a Fourier transform of the probability distribution in the environment along the direction perpendicular to the net relative shift induced by its coupling with the system. This leads back to sub-Planck scales: for small displacements the exponent in the integrand can be expanded. This yields

$$|\langle \epsilon_+ | \epsilon_- \rangle|^2 \approx 1 - \delta_p^2 (\langle x^2 \rangle - \langle x \rangle^2) / \hbar^2 \quad (17a)$$

The spatial extent of the wavefunction is naturally defined as $L = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$, where the averages of the observable x are over the state of the environment. The estimate of the shift δ_p leading to orthogonality is then

$$\delta_p \approx \hbar / \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \quad (17b)$$

in accord with the estimate of the dimensions of the sub-Planck structures, equation (8). The same simple formula holds when the environment is initially in a mixed state (see Methods).

Obvious generalizations of equations (16) and (17) are valid for arbitrary displacements, and for the case of many states of the system. This treatment can be also extended to interactions between the system and the environment that cannot be represented as simple shifts, although calculations become more complicated. The role played by the spread L of the state of the environment in the behaviour of the overlap (and hence in decoherence) is a direct consequence of the properties of the Fourier transform. Detailed analysis of other consequences of equation (16) is beyond the scope of this paper. Using elementary properties of the Fourier transform, readers can nevertheless confirm that the overlap decreases to near zero around δ_p (equations (8) and (17)), and that, for typical initial

states of the environment, it remains much less than unity. ‘Revivals’ of the overlap we have seen in the case of the compass state (Fig. 2) are now easily understood and dismissed as an exception. Equation (16) shows that they can happen only when the initial probability distribution of the environment in the direction perpendicular to displacement is localized to a few peaks, so that $|\epsilon(x)|^2$, the spectrum of the displacement-dependent overlap, is essentially discrete.

To conclude, I have shown that Wigner functions can, and generally will, develop phase space structures on scales as small as, but not generally smaller than, $a \sim \hbar^2/A$, equation (3). This is surprising, as sub-Planck scales are often regarded as unphysical. I have shown that they have remarkable physical implications: displacements given by $\delta \approx \hbar/\sqrt{A}$ suffice to induce orthogonality. They are a factor $\sim \sqrt{N} = \sqrt{(A/2\pi\hbar)}$ smaller than the displacement $\delta \approx \sqrt{\hbar}$ dictated by the Planck constant and needed to move a typical minimum-uncertainty gaussian in a random direction enough to noticeably reduce its overlap with its (old) self.

Nonlinear, and, especially, chaotic dynamics lead to states that quickly spread over much of the available phase space^{2,12–15}. Hence, one would expect that environments with unstable dynamics would be much more efficient decoherers, as they are constantly evolving into delocalized states. Yet standard models of decoherence^{17–19} use harmonic oscillators. They make up for this inefficiency by using many (infinity) of them, so that each becomes slightly displaced by the interaction with the system.

On a more mundane level, these results allow one to anticipate the structure of the mesh required to simulate evolution of a quantum system in phase space. They are related to the quantization of discrete chaotic maps on a torus²⁰, which turn out to require a mesh similar to the one given by equations (7) and (8). This is in contrast to simulations of classical systems, which are doomed by equation (1), as it implies resolution exponentially increasing with time. Most importantly, a controls the sensitivity of quantum states to perturbations, with obvious implications for decoherence outlined above. Moreover, sensitivity to perturbations will set limits on the ‘hypersensitivity’ of quantum chaotic systems²¹, and help one to understand the enhanced capacity of quantum chaotic systems for entanglement²².

The sensitivity of quantum systems in highly delocalized states may not just be a cause of accelerated decoherence (and hence an impediment to truly quantum applications) but in certain settings it may be beneficial. This is not as far-fetched as it may seem at first. After all, a detector in a compass-like state of Fig. 2 would be sensitive to perturbations $\sim \hbar/\sqrt{A}$ that are minute compared with the ‘standard quantum limit’²³. And, when the weak force to be detected perturbs its state enough to make it orthogonal to the initial state, a record that is in principle distinguishable has been made. The sensitivity of quantum meters would be then limited by a , in accord with equations (3), (16) and (17). Thus, it seems possible that the Schrödinger cat may eventually follow the path of other paradoxical quantum *gedanken* experiments that have found—after dusting off the classical preconceptions and shedding the aura of the paradox—far more useful employment in potential applications of quantum physics. $\square$

Methods

Decoherence happens to a quantum system S as a consequence of a measurement-like interaction with the environment E , which entangles their states:

$$|s\rangle|e\rangle = (\alpha|+\rangle + \beta|-\rangle)|e\rangle \rightarrow \alpha|+\rangle|\epsilon_+\rangle + \beta|-\rangle|\epsilon_-\rangle = |\Phi_{SE}\rangle$$

Above, we have assumed a system with a two-dimensional Hilbert space spanned by the $\{|+\rangle, |-\rangle\}$ orthonormal basis. The two conditional states of the environment $|\epsilon_\pm\rangle = U_\pm|e\rangle$ evolve under the unitary transformations U_+ (U_-) induced by the system in the state $|+\rangle$ ($|-\rangle$) respectively. When the effect of the interaction is simply a displacement in phase space, then $U_\pm = D_\pm = \exp(i(\delta_x^\pm p + \delta_p^\pm x)/\hbar)$, where D is the displacement operator, and $\delta^\pm = (\delta_x^\pm, \delta_p^\pm)$ are the resulting shifts. Such a displacement could be induced by a hamiltonian of interaction $H_{SE} = g(|+\rangle\langle+| - |-\rangle\langle-|)(\hbar/i)\partial_x$, where g is a

coupling constant. More general conditional evolutions of the environment $\mathcal{E}$ can be of course considered.

Following entanglement, the state of the system alone is described by the reduced density matrix obtained from $|\Phi_{SE}\rangle$ by a trace over the environment:

$$\rho_S = \text{Tr}_{\mathcal{E}} |\Phi_{SE}\rangle \langle \Phi_{SE}| = |\alpha|^2 |+\rangle \langle +| + z\alpha\beta^* |+\rangle \langle -| + z^*\alpha\beta^* |-\rangle \langle +| + |\beta|^2 |-\rangle \langle -|$$

Disappearance of the off-diagonal terms signifies perfect decoherence. In the $\{|+\rangle, |-\rangle\}$ basis this is guaranteed when the overlap $z = \langle \epsilon_+ | \epsilon_- \rangle = \text{Tr}(\epsilon_- \chi \epsilon_+)$ disappears. It is therefore natural to measure effectiveness of decoherence by the magnitude of the overlap of the two conditional states of the environment, which in turn determines the degree of suppression of the off-diagonal terms in ρ_S .

This two-paragraph 'crash course' is no substitute for a more complete discussion of decoherence⁶⁻⁸. We have swept a number of issues under the rug. Foremost among them is einselection—the emergence, in the course of decoherence, of the preferred set of pointer states that habitually appear on the diagonal of ρ_S essentially independently of the initial states of either $\mathcal{E}$ or $\mathcal{S}$. Stability of these pointer states (rather than the diagonality of the density matrix in some basis) is the key to the role played by decoherence in the transition from quantum to classical^{2-8,18}.

Another important subject avoided here is the likely situation when the state of the environment is represented by a mixture ρ_E . Detailed discussion of this case (treated extensively before^{6-8,17-19}, although not from the point of view of the sub-Planck structures) is beyond the scope of this paper, but the basic conclusion is easy to state: the estimated magnitude of the smallest displacement leading to orthogonality is still given by equation (17). That is, it is still related to the smallest scales compatible with the classical action A associated with ρ_E . Off-diagonal terms of ρ_S are suppressed by $z = \text{Tr} U_- \rho_E U_+^\dagger$, which is the relevant generalization of equation (16). This expression can be expanded for simple displacements $U_\pm = D_\pm$ in the limit of small shifts to recover equation (17), with the only difference arising from the fact that now the mixture ρ_E must be used to obtain the averages $\langle x^2 \rangle = \text{Tr} x^2 \rho_E$, etc.).

Note that the size of the structures in phase space depends on the representation of the quantum state. For instance, one does not expect scale a of equation (3) in the Husimi function, that is, in the Wigner function smoothed on a Planck scale. This is no surprise and does not affect our conclusions concerning decoherence, as only the Wigner function yields scalar product through the simple equation (14).

Note that the discussion throughout the paper is set in one spatial dimension. Generalization to d dimensions is as conceptually straightforward as it is notationally cumbersome. The structure saturates in volumes of a^d , etc. This has little effect on decoherence, as it depends on displacements that yield orthogonality, and these are still $\delta \approx \sqrt{a}$.

Received 25 September 2000; accepted 25 June 2001.

- Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Phys.* **43**, 172–198 (1927); The physical content of quantum kinematics and mechanics (Engl. Trans.) in *Quantum Theory and Measurement* (eds Wheeler, J. A. & Zurek, W. H.) (Princeton Univ. Press, Princeton, 1983).
- Zurek, W. H. Decoherence, chaos, quantum-classical correspondence, and the algorithmic arrow of time. *Phys. Scripta* **T76**, 186–198 (1998).
- Zurek, W. H. Pointer basis of a quantum apparatus: Into what mixture does the wavepacket collapse? *Phys. Rev. D* **24**, 1516–1524 (1981).
- Zurek, W. H. Environment-induced superselection rules. *Phys. Rev. D* **26**, 1862–1880 (1982).
- Joos, E. & Zeh, H. D. The emergence of classical properties through the interaction with the environment. *Z. Phys. B* **59**, 229 (1985).
- Zurek, W. H. Decoherence and the transition from quantum to classical. *Phys. Today* **44**, 36–46 (1991).
- Giulini, D., Joos, E., Kiefer, C., Kupsch, J., Stamatescu, L.-O. & Zeh, H. D. *Decoherence and the Appearance of a Classical World in Quantum Theory* (Springer, Berlin, 1996).
- Zurek, W. H. Decoherence, einselection, and the quantum origin of the classical. *Rev. Mod. Phys.* (in the press); also as preprint (quant-ph 010527) at (<http://xxx.lanl.gov>) (2001).
- Haake, F. *Quantum Signatures of Chaos* (Springer, Berlin, 1991).
- Casati, G. & Chirikov, B. *Quantum Chaos* (Cambridge Univ. Press, Cambridge, 1995).
- Hillery, M., O'Connell, R. F., Scully, M. O. & Wigner, E. P. Distribution functions in physics: Fundamentals. *Phys. Rep.* **106**, 121–167 (1984).
- Berry, M. V. & Balazs, N. L. Evolution of semiclassical quantum states in phase space. *J. Phys. A* **12**, 625–642 (1979).
- Korsch, H. J. & Berry, M. V. Evolution of Wigner's phase-space density under a nonintegrable quantum map. *Physica D* **3**, 627–636 (1981).
- Zurek, W. H. & Paz, J. P. Decoherence, chaos, and the Second Law. *Phys. Rev. Lett.* **72**, 2508–2511 (1994).
- Berman, G. P. & Zaslavsky, G. M. Condition of stochasticity in quantum non-linear systems. *Physica (Amsterdam)* **91A**, 450 (1978).
- Habib, S., Shizume, K. & Zurek, W. H. Decoherence, chaos, and the correspondence principle. *Phys. Rev. Lett.* **80**, 4361 (1998).
- Caldeira, A. O. & Leggett, A. J. Path-integral approach to quantum Brownian motion. *Physica* **121A**, 587–616 (1983).
- Paz, J. P. & Zurek, W. H. in *Les Houches Lectures Session LXXXII* (eds Kaiser, R., Westbrook, C. and David, F.) 533–614 (Springer, Berlin, 2001).
- Braun, D., Haake, F. & Strunz, W. A. Universality of decoherence. *Phys. Rev. Lett.* **86**, 2913–2917 (2001).
- Hannay, J. H. & Berry, M. V. Quantization of linear maps on a torus—Fresnel diffraction by a periodic grating. *Physica* **1D**, 267–290 (1980).
- Caves, C. in *Physical Origins of Time Asymmetry* (eds Halliwell, J. J., Pérez-Mercader, J. & Zurek, W. H.) 47–77 (Cambridge Univ. Press, Cambridge, 1993).
- Miller, P. A. & Sarkar, S. Signatures of chaos in the entanglement of two coupled quantum kicked tops. *Phys. Rev. E* **60**, 1542 (1999).

- Braginsky, V. B. & Khalili, F. Y. Quantum nondemolition measurements: the route from toys to tools. *Rev. Mod. Phys.* **95**, 703–711 (1996).
- Karkuszewski, Z., Zakrzewski, J. & Zurek, W. H. Breakdown of correspondence in chaotic systems: Ehrenfest versus localization times. Preprint quant-ph/0010011 at (<http://xxx.lanl.gov>) (2000).

Acknowledgements

This research was supported in part by the National Security Agency. I thank A. Albrecht, N. Balazs, C. Jarzynski, Z. Karkuszewski and J. P. Paz for useful chaotic conversations.

Correspondence and requests for materials should be addressed to W.Z. (e-mail: whz@LANL.gov).

Crystalline ion beams

T. Schätz, U. Schramm & D. Habs

Sektion Physik, LMU München, D-85748 Garching, Germany

By freezing out the motion between particles in a high-energy storage ring, it should be possible¹⁻⁴ to create threads of ions, offering research opportunities beyond the realm of standard accelerator physics. The usual heating due to intra-beam collisions should completely vanish, giving rise to a state of unprecedented brilliance. Despite a continuous improvement of beam cooling techniques, such as electron cooling and laser cooling, the ultimate goal⁵ of beam crystallization has not yet been reached in high-energy storage rings. Electron-cooled dilute beams of highly charged ions show liquid-like order^{6,7} with unique applications⁸. An experiment⁵ using laser cooling^{9,10} suggested a reduction of intra-beam heating, although the results were ambiguous. Here we demonstrate the crystallization of laser-cooled Mg^{+} beams circulating in the radiofrequency quadrupole storage ring PALLAS^{11,12} at a velocity of $2,800 \text{ m s}^{-1}$, which corresponds to a beam energy of 1 eV. A sudden collapse of the transverse beam size and the low longitudinal velocity spread clearly indicate the phase transition. The continuous ring-shaped crystalline beam shows exceptional stability, surviving for more than 3,000 revolutions without cooling.

In ion storage rings¹³, gaseous ion beams are heated through scattering within the beam. In combination with the varying focusing and bending elements, this mechanism couples part of the beam energy into the random ion motion. Cooling increases the phase-space density of the beam, and further amplifies the scattering rate¹⁴. In high-energy rings, this vicious circle can only be overcome by increasing the number of focusing sections¹⁵, or, in practice, by choosing a sufficiently low density of stored ions⁵. Recently, space-charge-limited densities have been reached¹⁰, but no beam crystallization has been observed so far, in striking contrast to the routine generation of elongated ion crystals at rest in linear^{16,17} and ring traps^{12,18}. As an illustration, in Fig. 1 we present images of ion crystals at rest¹², gained with our storage ring PALLAS (Paul laser cooling acceleration system), described below. Ions remain ordered because their mutual Coulomb repulsion overcomes their mean kinetic energy. The overall Coulomb repulsion is compensated by an external parabolic trapping potential ψ . The formation of the crystalline structure is well understood^{4,17-20}. It develops from a linear chain of ions over a zig-zag band to three-dimensional helices when the linear ion density¹⁹ λ is increased, either by adding more ions or by reducing the confining potential.

The difference between the behaviour in storage rings and traps might be caused by the predicted excitation of ion crystals passing through the periodic bending and focusing sections of a ring²⁰⁻²². To bridge this gap and experimentally elucidate the conditions necessary for obtaining crystalline beams in high-energy rings, we

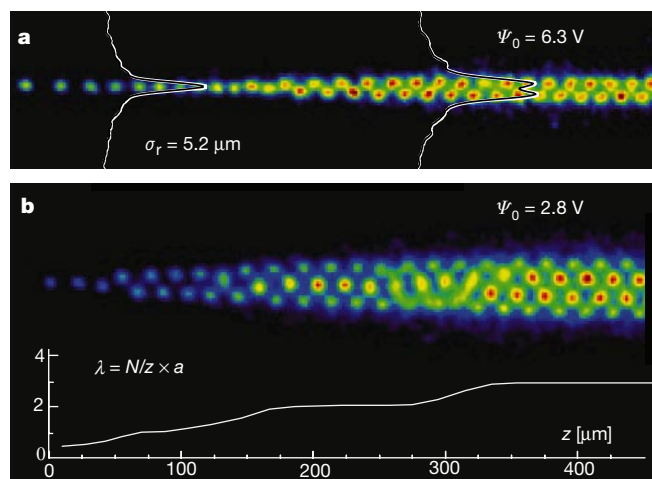


Figure 1 Images of ion crystals at rest in PALLAS. False colours reflect the fluorescence intensity of individual ions. The ions are longitudinally confined in a weak static potential¹² which is generated by the two drift tubes, highlighted in Fig. 2. The crystal becomes more complex when the linear ion density $\lambda = (N/z) \times a$, N denoting the number of particles and a the Wigner–Seitz radius¹⁹, increases with lowered confining potential ψ_0 from **a** to **b** or when it increases stepwise along the axis z because of the weak longitudinal confinement, as illustrated in **b**. The radius $\sigma_r = 5.2 \mu\text{m}$ of the linear ion string mainly reflects the overall spatial resolution (integration time 0.4 s).

constructed the low-energy radio frequency (r.f.) quadrupole storage ring PALLAS¹¹, which resembles a linear Paul trap, bent in a circle¹⁸ (see Fig. 2). Sixteen individual drift tubes enclose the quadrupole rods and can be used to transport and position ions along the orbit¹². Care has been taken to keep potential distortions on axis below 100 meV and to provide perfect grounding of the drift tubes to facilitate ion acceleration and to reduce heating due to longitudinal excitations. Radial confinement is achieved by applying an r.f. voltage $U_0 \cos(\Omega t)$ of a typical amplitude $U_0 = 200 \text{ V}$ and frequency $\Omega = 2\pi \times 6.3 \text{ MHz}$ between the quadrupole ring electrodes. Stored ions experience a harmonic pseudo-potential $\psi(r) = \psi_0 r^2/r_0^2$ with a depth of $\psi_0 = eU_0^2/4m\Omega^2 r_0^2 = 4 \text{ V}$, where e and m stand for the ion charge and mass of the $^{24}\text{Mg}^+$ ions and

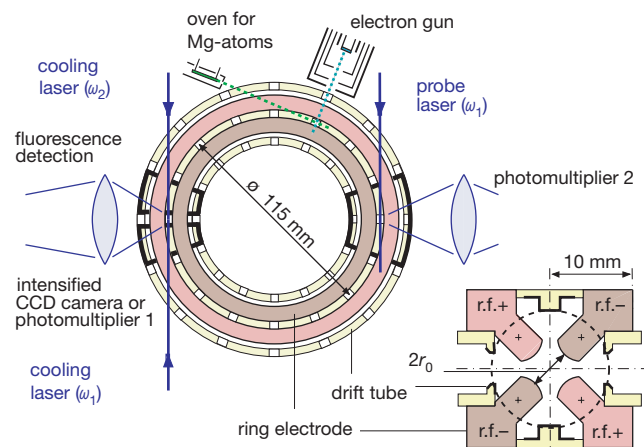


Figure 2 Axial and radial cut through the r.f. quadrupole storage ring PALLAS. The length of the orbit enclosed by the ring-shaped quadrupole electrodes is 0.36 m. Sixteen drift tubes are distributed around the ring, which can be powered individually. The drift tubes at the two opposite locations for laser cooling and laser probing are highlighted. To load the ring, ^{24}Mg atoms are ionized inside the trapping volume by electron bombardment.

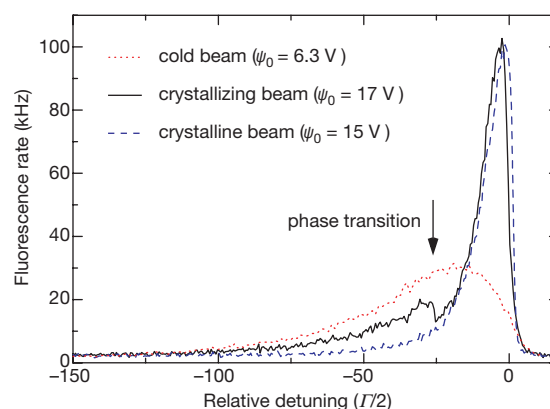


Figure 3 Fluorescence rate of the ion beam as a function of the frequency detuning of the co-propagating laser. The detuning is given in terms of half the natural transition linewidth ($\Gamma = 2\pi \times 42.7 \text{ MHz}$). The laser is slowly (4 s) tuned towards the resonance with the counter-propagating laser at fixed frequency, which defines the nominal ion velocity of $2,800 \text{ m s}^{-1}$. The beam contained 18,000 $^{24}\text{Mg}^+$ ions, as deduced from the fluorescence rate which was calibrated against the ion crystal at rest. The marked discontinuity in the solid curve (arrow) is characteristic of the phase transition to the crystalline beam.

$r_0 = 2.5 \text{ mm}$ for the aperture radius. The oscillation of ions in this confining pseudo-potential corresponds to the betatron oscillation¹³ of ions in a high-energy ring, and the period of the time-varying potential corresponds to its discrete focusing sections. For the given parameters, the number of betatron oscillations per round trip Q , which characterizes the focusing conditions, amounts to $Q = 50$. The ion velocity spread is reduced by means of standard laser cooling, exploiting the Doppler-shifted $3s^2S_{1/2} - 3p^2P_{3/2}$ transition of the moving $^{24}\text{Mg}^+$ ions.

Ions are accelerated by the light pressure of the continuously tuned co-propagating laser (force $+F_1$). The final beam velocity of $v = 2,800 \text{ m s}^{-1}$ is defined by the counter-propagating and thus decelerating laser ($-F_2$), kept at fixed frequency. The longitudinal velocity spread of the beam is reduced by the dispersive character of the combined laser force ($F_1 - F_2$). The typical behaviour of a circulating cold ion beam is shown by the dotted curve in Fig. 3. As the detuning of the co-propagating laser is decreased, the fluorescence rate increases. Then, as it approaches the resonance, the forces start to compensate and the rate drops off again. Increasing the confining potential markedly changes the behaviour (solid curve). Here the fluorescence at first follows the previous curve, then decreases abruptly, and subsequently rises to a sharp peak whose width is dominated by the saturation-broadened linewidth of the transition. This signature of an abrupt decrease has been previously observed in ion traps²³ and indicates the phase transition to the crystalline state. After a slight reduction of the potential, the phase transition can no longer be resolved (dashed curve). This behaviour suggests that the ions are confined strongly enough to form the string but do not experience unnecessary r.f. heating, which is responsible for the visibility of the phase transition.

Additionally, the phase transition is pinpointed by a sudden decrease of the transverse beam size (Fig. 4a, b). A comparison of the spatial profile with that in Fig. 1a specifies the formation of a circulating ion string. The number of particles in the string (18,000) was deduced from the overall fluorescence yield and denotes a mean interparticle distance of $d \approx 20 \mu\text{m}$ ($\lambda = 0.4$). The beam radius σ_r corresponds to an initial transverse temperature of $T_{\perp} = 30 \text{ K}$ for the gaseous and $T_{\perp} < 0.4 \text{ K}$ (resolution-limited) for the crystalline beam. The longitudinal velocity spread of the beam was probed opposite to the cooling section (see Fig. 2). The line profiles in Fig. 4c and d reveal a width of the velocity distribution of 4 m s^{-1} ($T_{\parallel} = 50 \text{ mK}$) for the cold gaseous beam and less than 1 m s^{-1}

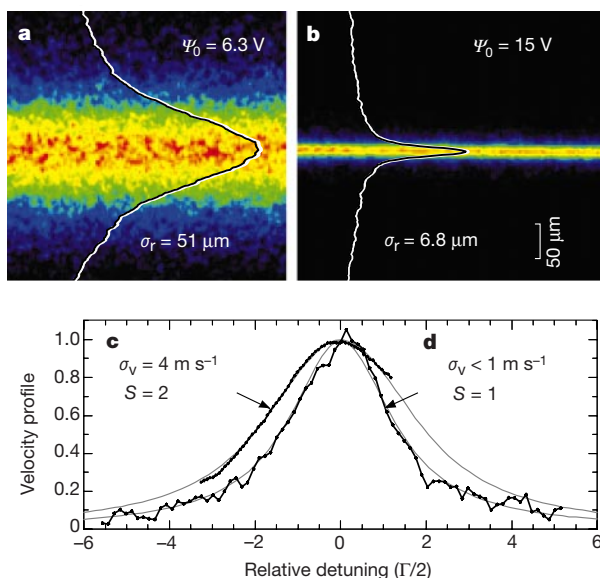


Figure 4 Transverse beam profiles and velocity profiles before (a, c) and after (b, d) the phase transition. The transition is induced by increasing the radial focusing strength Ψ_0 , forcing the ions, already sufficiently cold, into the linear string. To prove the longitudinal velocity distribution, the ions are periodically velocity shifted by the pair of drift tubes shown in Fig. 2 (amplitude ~ 60 mV, frequency 1 Hz). By use of the Doppler effect they are tuned over the resonance with the fixed frequency probe laser and the fluorescence rate is recorded. Voigt-profiles (grey lines) were fitted to disentangle the Doppler contribution σ_v for the two measurements from the saturation-broadened width of the transition (S denotes the saturation).

($T_{\parallel} < 3$ mK) for the crystalline beam. The inter-ion coupling strength is usually described by the plasma parameter $\Gamma_p = (e^2/4\pi\epsilon_0 d)(1/kT)$, the ratio of the mutual potential energy of ions to their mean thermal energy kT . In the longitudinal direction, Γ_p is 250 for the circulating ion string. It even exceeds the theoretical threshold of 180 which is required for the formation of three-dimensional ion crystals^{20,24}. The independent measurement of the transverse temperature of the crystal reveals a value of $\Gamma_p > 2$, consistent with the longitudinal value and pointing to the strongly coupled regime¹⁵, where both components have to be identical. So the full phase-space density has been determined to increase by a factor of at least 10^4 and, assuming $T_{\perp} = T_{\parallel}$, beyond 10^6 in the phase transition.

One of the outstanding properties of ion crystals is their elasticity which leads to a strong suppression of the coupling of the periodic r.f. motion into random thermal motion²⁵. Under continuous laser cooling, circulating ion strings were observed to survive for many hours without any significant ion loss. For the study of heating rates, both cooling lasers were simultaneously blocked for a given period of time. The lasers were unblocked and the remaining fluorescence rate was measured. For the ion crystals at rest (Fig. 1b), the fluorescence rate immediately reappears at its full strength for blocking periods of up to 90 s (Fig. 5, upper frame). This indicates the persistence of the crystalline state. In contrast, after a blocking time of 250 s the fluorescence rate is characteristic for the gaseous cloud. The survival time of 6×10^6 r.f. periods reasonably coincides with simulations²⁵ of ion crystal stability in time-varying potentials. A similar procedure for the crystalline ion beam also results in an immediate restoration of the fluorescence signal after 40 ms of blocking the lasers (Fig. 5d). After a period of 400 ms (Fig. 5e) the signal reappears at a level of 50%, a signal rate that reflects the survival of a less dense crystalline beam. Moreover, the signal fully recovers on a millisecond timescale. We therefore suggest that a two-phase regime is formed, in which a string of decreased ion number

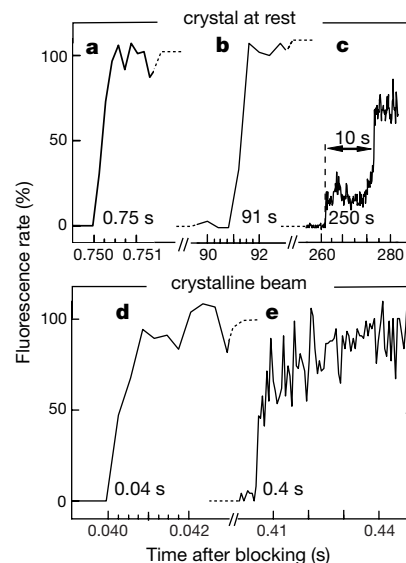


Figure 5 Fluorescence signal of the ion crystal at rest and the crystalline beam after blocking and unblocking the cooling lasers. The crystal at rest (a–c) survives blocking periods of the order of seconds to minutes, where for a the time resolution of 250 μ s excludes the possibility of melting and subsequent re-crystallization within the rise-time of the fluorescence signal. For the blocking period of 250 s the crystal melts (c) and re-crystallizes after 10 s of cooling on a timescale of seconds. The crystalline beam shows an instantaneous recovery after 40 ms (d). After blocking the cooling for 400 ms (e), the count rate jumps back to 50% of the initial level and fully recovers within 10 ms.

survives and a cold beam halo has to be rearranged into the crystalline beam.

The realization of the circulating ion string and the prospect of attaining vertical zig-zag and helical structures in PALLAS should allow experimental investigation of many questions concerning the formation of ion crystals in high-energy rings, currently known only from simulations. The problem of shear and the potential necessity of cooling to constant angular velocity, as well as the complex interplay between the ion crystal and the periodic focusing of the ring^{21,22}, can be addressed in PALLAS and directly transferred to large-scale rings¹¹. The drift tubes can be used to alter the focusing conditions and the velocity can at present be varied between 10^5 and 10^4 m s^{−1}. Bunching of an ion beam, one important feature of high-energy rings, has also been studied and will be reported elsewhere. □

Received 20 April; accepted 24 June 2001.

1. Parkhomchuk, V. in *Crystalline Beams and Related Issues* (eds Maletic, D. M. & Ruggiero, A. G.) 409–420 (World Scientific, Singapore, 1996).
2. Pestrikov, D. V. in *Crystalline Beams and Related Issues* (eds Maletic, D. M. & Ruggiero, A. G.) 275–294 (World Scientific, Singapore, 1996).
3. Schiffer, J. P. & Kienle, P. Could there be an ordered condensed state in beams of fully stripped heavy ions. *Z. Phys. A* **321**, 181 (1985).
4. Rahman, A. & Schiffer, J. P. Structure of a one-component plasma in an external field. A molecular-dynamics study of particle arrangement in a heavy-ion storage ring. *Phys. Rev. Lett.* **57**, 1133–1136 (1986).
5. Eisenbarth, U. et al. Anomalous behaviour of laser-cooled fast ion beams. *Hyperfine Interact.* **127**, 223–235 (2000).
6. Steck, M. et al. Anomalous temperature reduction of electron-cooled heavy ion beams in the storage ring ESR. *Phys. Rev. Lett.* **77**, 3803–3806 (1996).
7. Hasse, R. W. Theoretical verification of Coulomb order of ions in a storage ring. *Phys. Rev. Lett.* **83**, 3430–3433 (1999).
8. Radon, T. et al. Schottky mass measurements of stored and cooled neutron-deficient projectile fragments in the element range of 57. *Nucl. Phys. A* **677**, 75–99 (2000).
9. Lauer, I. et al. Transverse laser cooling of a fast stored ion beam through dispersive coupling. *Phys. Rev. Lett.* **81**, 2052–2055 (1998).
10. Madsen, N. et al. Density limitations in a stored laser-cooled ion beam. *Phys. Rev. Lett.* **83**, 4301–4304 (1999).
11. Schätz, T., Habs, D., Podlech, C., Wei, J. & Schramm, U. Towards crystalline ion beams—the PALLAS

- ring trap. in *Proc. Workshop on Trapped Charged Particles and Fundamental Physics* (eds Dubin, D. & Schneider, D.) *AIP Conf. Proc.* **457**, 269–273 (1999).
12. Schramm, U., Schätz, T. & Habs, D. in *Proc. Conf. on Appl. of Acc. in Research and Industry* (eds Duggan, J. L.) *AIP Conf. Proc.* **576** (in the press).
 13. Bryant, P. J. & Johnson, K. *Circular Accelerators and Storage Rings* (Cambridge Univ. Press, Cambridge, 1993).
 14. Spreiter, Q., Seurer, M. & Toepffer, C. Relaxation in a strongly coupled particle beam. *Nucl. Instrum. Methods A* **364**, 239–242 (1995).
 15. Seurer, M., Spreiter, Q. & Toepffer, C. in *Crystalline Beams and Related Issues* (eds Maletic, D. M. & Ruggiero, A. G.) 311–328 (World Scientific, Singapore, 1996).
 16. Raizen, M. G. *et al.* Ionic crystals in a linear Paul trap. *Phys. Rev. A* **45**, 6493–6501 (1992).
 17. Drewsen, M., Brodersen, C., Hornekaer, L., Hangst, J. S. & Schiffer, J. P. Large ion crystals in a linear Paul trap. *Phys. Rev. Lett.* **81**, 2878–2881 (1998).
 18. Birk, G., Kassner, S. & Walther, H. Multiple-shell structures of laser-cooled Mg-ions in a quadrupole storage ring. *Nature* **357**, 310–313 (1992).
 19. Hasse, R. W. & Schiffer, J. P. The structure of the cylindrically confined Coulomb lattice. *Ann. Phys.* **203**, 419–448 (1990).
 20. Habs, D. & Grimm, R. Crystalline ion beams. *Ann. Rev. Nucl. Part. Sci.* **45**, 391–428 (1995).
 21. Schiffer, J. P. in *Crystalline Beams and Related Issues* (eds Maletic, D. M. & Ruggiero, A. G.) 217–228 (World Scientific, Singapore, 1996).
 22. Wei, J., Okamoto, H. & Sessler, A. M. Necessary conditions for attaining a crystalline beam. *Phys. Rev. Lett.* **80**, 2606–2609 (1998).
 23. Blümel, R. *et al.* Phase transitions of stored laser-cooled ions. *Nature* **334**, 309–313 (1988).
 24. Dubin, D. H. E. First-order anharmonic correction to the free energy of a Coulomb-crystal in periodic boundary conditions. *Phys. Rev. A* **42**, 4972–4982 (1990).
 25. Schiffer, J. P., Drewsen, M., Hangst, J. & Hornekaer, L. Temperature, ordering, and equilibrium with time-dependent forces. *Proc. Natl Acad. Sci. USA* **97**, 10697–10700 (2000).

Acknowledgements

We thank R. Neugart for technical support, and P. Kienle and H. Walther for discussions. The work was partially funded by the Deutsche Forschungsgemeinschaft and the Maier Leibnitz Labor.

Correspondence and requests for material should be addressed to U.S.
(e-mail: ulrich.schramm@physik.uni-muenchen.de).

A titanosilicate molecular sieve with adjustable pores for size-selective adsorption of molecules

Steven M. Kuznicki*, Valerie A. Bell*, Sankar Nair†, Hugh W. Hillhouse†, Richard M. Jacobinas*, Carola M. Braunbarth†, Brian H. Toby‡ & Michael Tsapatsis†

* Strategic Technology Group, Engelhard Corporation, 101 Wood Avenue, Iselin, New Jersey 08830, USA

† Department of Chemical Engineering, 159 Goessmann Laboratory, University of Massachusetts, Amherst, Massachusetts 01003, USA

‡ NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8562, USA

Zeolites and related crystalline microporous oxides—tetrahedrally coordinated atoms covalently linked into a porous framework—are of interest for applications ranging from catalysis to adsorption and ion-exchange¹. In some of these materials (such as zeolite rho) adsorbates², ion-exchange, and dehydration and cation relocation^{3,4} can induce strong framework deformations. Similar framework flexibility has to date not been seen in mixed octahedral/tetrahedral microporous framework materials, a newer and rapidly expanding class of molecular sieves^{5–16}. Here we show that the framework of the titanium silicate ETS-4, the first member of this class of materials⁸, can be systematically contracted through dehydration at elevated temperatures to ‘tune’ the effective size of the pores giving access to the interior of the crystal. We show that this so-called ‘molecular gate’ effect can be used to tailor the adsorption properties of the materials to give size-selective adsorbents¹⁷ suitable for commercially important separations of gas mixtures of molecules with similar size in the 4.0 to 3.0 Å range, such as that of N₂/CH₄, Ar/O₂ and N₂/O₂.

ETS-4 has a mixed octahedral/tetrahedral framework (as shown in the diagram in Fig. 1) with a structure related to the mineral zorite^{18–20}, and has been described in terms of the random intergrowth of four hypothetical polymorphs²¹. Although larger openings are present in its structure, faulting ensures that access to the crystal interior of ETS-4 occurs through the relatively narrow eight-membered rings (8MRs), analogous to what is seen in small-pore zeolites²². The framework structure and cation positions of as-synthesized ETS-4 (Na-ETS-4) and of Sr-exchanged ETS-4 have been recently reported^{21,23}. A distinct feature of ETS-4 is the presence of titania semioctahedra that are connected to framework silicon atoms through only four oxygen bridges^{18,21}, which gives rise to a planar connectivity, in contrast to the three-dimensional connectivity commonly encountered in other microporous frameworks.

As-synthesized ETS-4 has been reported to collapse near 200 °C to an amorphous material owing to the loss of structural water chains present along the channel system¹⁵. We found that carefully controlled heating leads to a monotonic decrease of the lattice dimensions in all three crystallographic directions, along with the gradual loss of crystallinity. Upon appropriate ion-exchange (for example, with Sr) the thermal stability of the structure, as inferred from diffraction and water adsorption measurements, can be extended to higher temperatures (for example, up to 350 °C for Sr-exchanged ETS-4).

Water-loss curves (as determined by thermal gravimetric analysis, TGA) for as-synthesized and Sr-exchanged ETS-4, along with representative room-temperature powder neutron diffraction patterns of Sr-exchanged ETS-4 after dehydration at temperatures up to 300 °C, are given in Fig. 2a and b, respectively. It is evident that the exchange by Sr leads to higher temperatures for dehydration, and as a result extends the stability range of the ETS-4 framework. The lattice constants, determined from the room-temperature neutron diffraction data, are listed in the Fig. 2 legend, while the evolution of these constants, as determined by *in situ* X-ray diffraction (XRD), is shown in Fig. 2c. A sharp structural transition takes place at ~110 °C, followed by a continuous contraction up to the point when the material becomes amorphous (at temperatures higher than 350 °C). We refer to the dehydrated materials as CTS (contracted titanosilicates). Samples of Sr-exchanged ETS-4 treated at temperatures higher than 250 °C exhibit increasing loss of order, as shown in the powder X-ray and neutron diffraction patterns. Reflections with a *k* component broaden and disappear first, indicating deformation and eventual breaking of the titania chains that run along the *b* axis.

The contraction of the CTS samples is preserved on cooling to room temperature in a dry environment. Samples treated at a temperature below 250 °C and rehydrated by exposure to moist air recover their original structure, as determined by XRD. The reversibility of the contraction/expansion phenomenon of ETS-4 points to the well known role of cation relocation on loss of water in inducing framework distortions^{24–26}. Samples dehydrated at higher temperatures and exhibiting partial loss of X-ray crystallinity rehydrate more slowly, and do not fully recover their original crystalline structure upon rehydration. The unit cell dimensions of the partially disordered material produced by dehydration of Sr-exchanged ETS-4 at temperatures higher than 330 °C are stable upon exposure to moist air for at least two weeks. The irreversible change to the partially disordered structure (CTS) can be attributed to the destruction of connectivity of the titania chains and consequent loss of the restoring force needed to recover the original structure upon rehydration; strong cation coordination to framework oxygen atoms also promotes this irreversible change. Heat treatment at slightly higher temperatures (>350 °C) leads to complete loss of crystalline order.

Rietveld refinement^{27–29} of the powder neutron diffraction data of Fig. 2b revealed a progressive and pronounced effect on the 8MR-opening in dehydrated samples (Fig. 3), in addition to cation

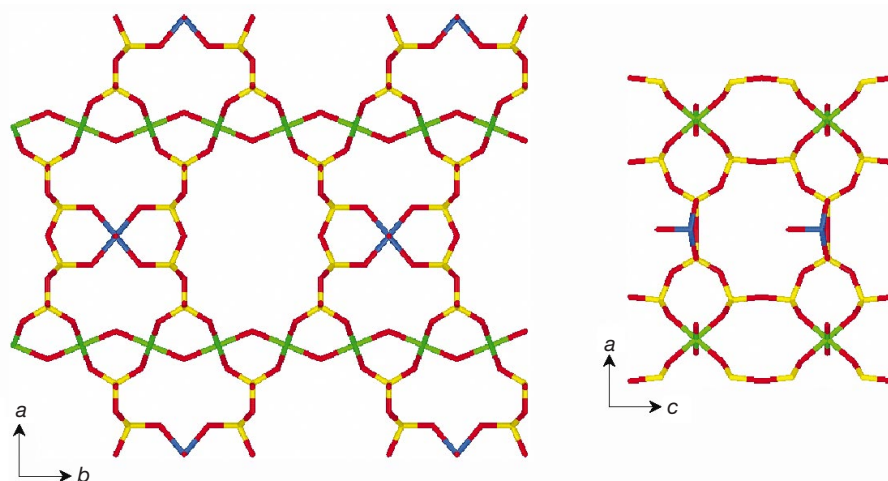


Figure 1 ETS-4 framework. Octahedrally coordinated titanium atoms that form O–Ti–O–Ti–O– chains along the *b* direction are shown in green. The titania semioctahedra of the bridging units that are connected to four framework silicon atoms through oxygen bridges are shown in blue. Silicon and oxygen atoms are shown in yellow and red, respectively. The *a*–*b* slice (left) shows the 12MR opening while the *a*–*c* slice (right) shows the 8MR

opening. Pore connectivity along the *a* direction is through 6MRs. As a result, transport of molecules in ETS-4 can be considered to be two-dimensional within the *b*–*c* plane channel system. Owing to faulting, the connectivity of the 12MR down the *c* axis is interrupted and, therefore, transport and access to the interior of ETS-4 is controlled by the 8MR opening.

relocation. More specifically, with the contraction of the unit cell dimensions of ETS-4 during dehydration, the 8MR opening becomes asymmetric, with the D_1 distance (Fig. 3) decreasing monotonically. At the same time, cation relocation takes place and two new cation positions appear. One of the new cation

positions (Sr3; see Supplementary Information Table I) is at the centre of the 8MR and its occupancy increases with the temperature of dehydration. Although this new cation position has a low occupancy, it nevertheless has pronounced effects on the access to the crystal interior owing to the two-dimensional connectivity of

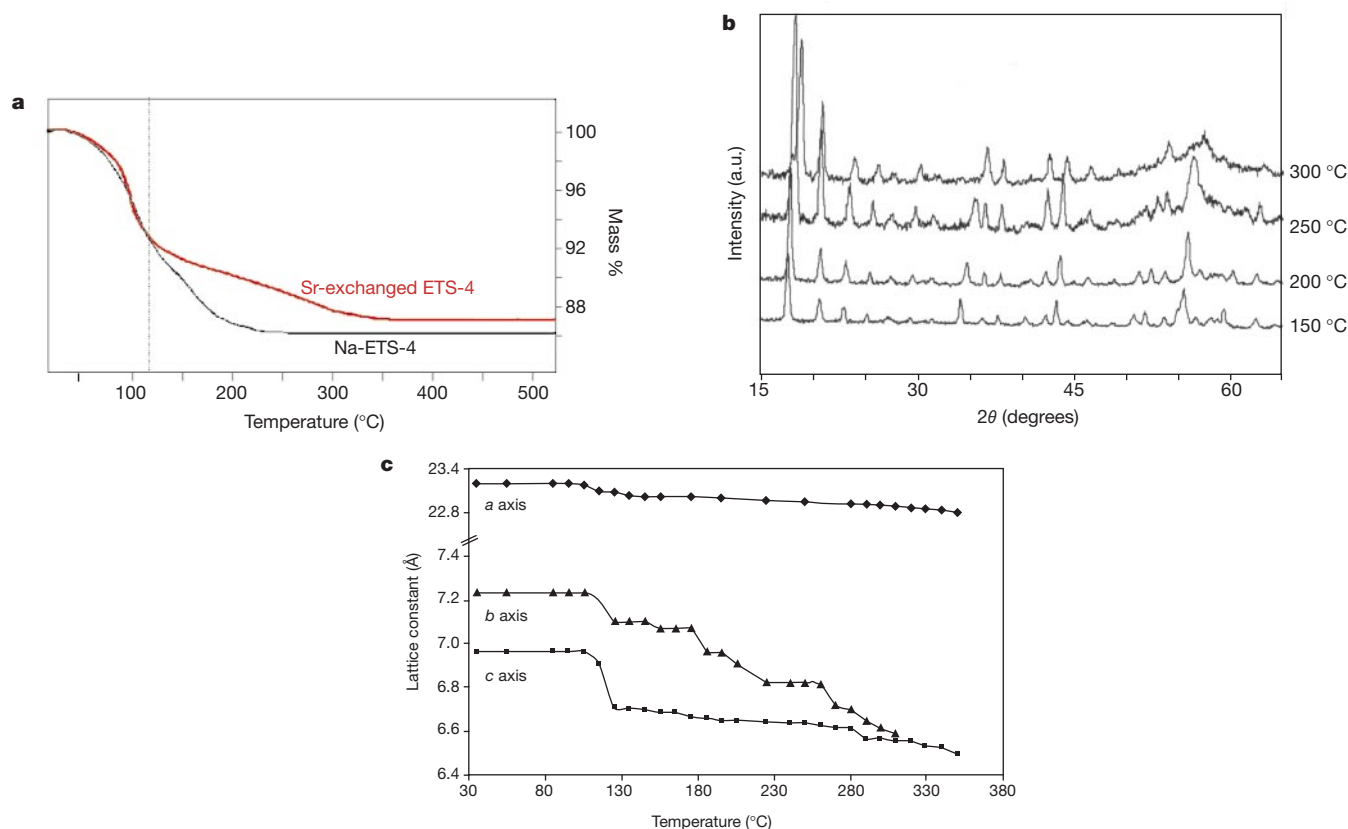


Figure 2 ETS-4 dehydration and associated decrease of the lattice constants. **a**, Thermogravimetric analysis (TGA) showing differences in the water loss of as-synthesized ETS-4 (Na-ETS-4) and Sr-exchanged ETS-4. **b**, Powder neutron diffraction patterns of Sr-exchanged ETS-4 collected at room temperature after dehydration at 150, 200, 250 and 300 °C. The corresponding lattice constants (*a* axis, *b* axis, *c* axis in Å) are:

(23.0120, 7.0705, 6.6899), (23.0006, 6.9588, 6.6486), (22.9465, 6.8198, 6.6394), (22.9007, 6.6171, 6.5550) while the Le Bail weighted residuals (R_{wp}) are 1.68, 3.09, 3.41 and 3.56, respectively. The lattice parameters for the as-made sample are 23.1962, 7.2381, 6.965221. **c**, Changes of the lattice constants of Sr-exchanged ETS-4 during heat treatment as determined by *in situ* XRD.

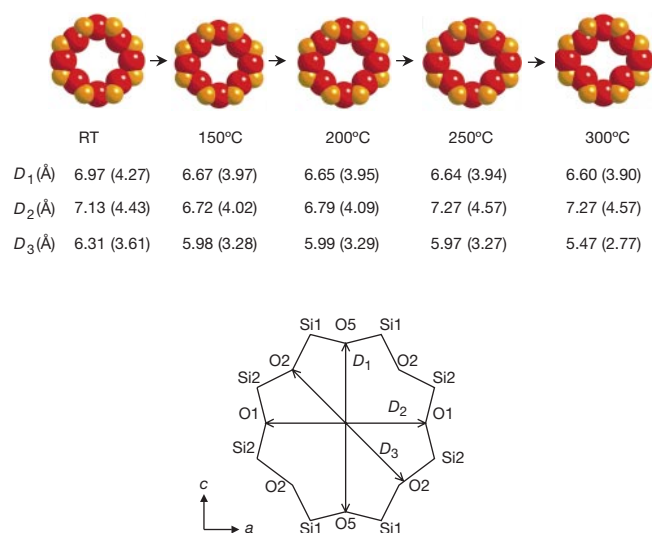


Figure 3 Reduction in the eight-membered ring (8MR) pore opening with increasing dehydration temperature. Oxygen to oxygen distances in the 8MR (D_1 , D_2 and D_3) for the room-temperature structure of Sr-exchanged ETS-4 and for the samples dehydrated at 150, 200, 250 and 300 °C were calculated on the basis of atomic positions obtained from Rietveld refinements (see ref. 21, and Supplementary Information Table 1a–d and Fig. 1a–d). Oxygen atoms with identical labels are related by mirror planes. The O5–O5 (D_1) distance is equal to the lattice constant along c (see Fig. 2). The numbers in brackets are the effective distances obtained by subtracting twice the oxygen radius (1.35 Å; refs 30, 32) from the distances between the centres of the atoms. For a discussion of the size of the oxygen in molecular sieves, see ref. 32 and references therein. RT, room temperature.

the channel system of ETS-4 (Fig. 1). Both pore narrowing and cation relocation lead to a progressive reduction of the effective diameter of the 8MR.

The progressive contraction of the effective pore size of the 8MR opening profoundly affects the adsorption properties of these materials. Figure 4 shows a clear correlation between a measure of the 8MR opening and the material's adsorption capacity for various gas molecules. We note that although structure refinement of Sr-exchanged ETS-4 dehydrated at temperatures higher than 300 °C was not attempted owing to loss of crystallinity, the D_1 distance (see Fig. 3) of the 8MR opening coincides with the (001) d -spacing and can be reliably calculated from X-ray diffraction data. From Figs 2c, 3 and 4 it is evident that as dehydration temperature is increased and effective pore size³⁰ declines, molecules of progressively smaller dimensions are excluded from adsorption into the crystals.

This phenomenon, the contraction of the unit cell, is cation dependent. It occurs over the temperature regime associated with the loss of those crystal water molecules that are coordinated to the extra-framework cations. Other cation forms than those reported here show similar behaviour (for example, Ba-exchanged ETS-4); however, the specific cation or cation combination has a substantial effect on the temperature range over which shrinkage occurs. Partial exchange with Sr leads to similar behaviour, but the temperature range that leads to framework contraction through dehydration is larger. Figure 5 illustrates size-dependent separation of gas molecules using partially (75%) Sr-exchanged ETS-4 whose 8MR pores have been progressively contracted through dehydration. When the material is calcined at 190 °C, methane is readily adsorbed while the larger ethane molecules are essentially excluded. The capacity for methane absorption declines with further pore contraction; once the material has dehydrated at 270 °C, substantial exclusion occurs whereas the smaller nitrogen molecules are readily absorbed. Dehydration at 300 °C results in sufficiently pronounced pore contraction to show signs of nitrogen exclusion while smaller

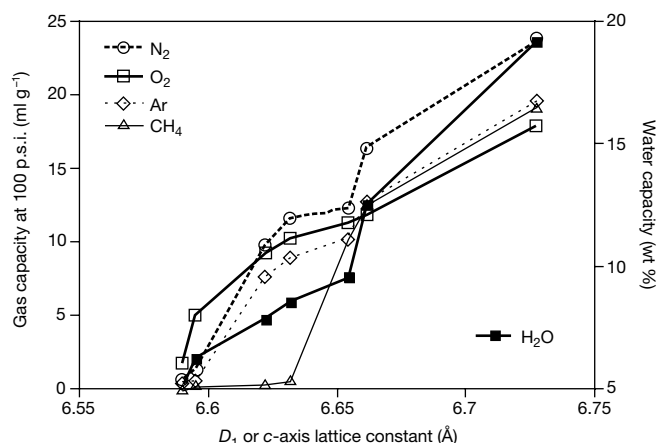


Figure 4 Sorption of molecules of various sizes with lattice contraction of Sr-exchanged ETS-4. This is the material used to provide the data in Figs 2 and 3; adsorption was performed at 25 °C for an equilibration time of 1 hour. 10-hour and 1-hour equilibrations yielded essentially identical results. The dehydration temperatures used to obtain the given D_1 values range from 150 to 350 °C. The corresponding Pauling dimensions³⁰ are: for CH₄, 4.2 Å; Ar, 3.84 Å; N₂, 4.1 Å (length), 3.00 Å (width); O₂, 3.9 Å (length), 2.8 Å (width); and H₂O, 3.9 Å (length), 3.15 Å (width).

oxygen molecules can still penetrate the crystal, resulting in an oxygen-selective adsorbent. The dominant cause of loss of gas capacity for a species able to enter the crystal is reduction in crystallinity, with a lesser effect due to lattice shrinkage.

We refer to the ability to systematically manipulate the effective pore size in these materials as the 'molecular gate' effect, which enables the separation of molecules of nearly identical size. An important example is the separation of nitrogen from methane. Nitrogen occurs commonly in natural gas, to the extent that its presence can render many natural-gas reservoirs unusable. As Figs 4 and 5 illustrate, the effective ETS-4 pore size can be manipulated to exclude methane while still allowing the adsorption of the smaller nitrogen molecule. First field demonstrations show that the molecular gate phenomenon allows removal of nitrogen from natural gas (with water vapour contents of 80–150 p.p.m.) at well-head pressures, reducing an initial nitrogen content of 18% to less than 5% with a methane recovery of at least 90%³¹. The system consists of seven pressure-swing adsorption beds with overall process cycling time of 350 seconds, or 50 seconds adsorption per bed. On a laboratory scale, many other separations, including the major constituents of air (Ar/O₂ and N₂/O₂), have been accomplished¹⁷.

The adsorption and structural results point to the contraction of the 8MRs and concomitant cation relocation as being primarily responsible for the gradual exclusion of smaller and smaller molecules. However, some contributions may be due to the presence and potential pore-blocking role of extra-framework species, which may result from the progressive loss of crystalline order that accompanies pore contraction. We have industrially manufactured ETS-4 on a multi-ton scale, using standard equipment and raw materials common to the zeolite industry. Despite the unit-cell contraction and disorder induced by dehydration, it retains adsorption capacities and uptake rates suitable for practical applications. For example, the adsorption step in removing nitrogen from natural gas can be completed in less than one minute. The kinetic stability of ETS-4 to rehydration allows the material to be used in relatively humid process streams. We have a pilot plant that has successfully purified methane using raw, wet natural gas for nearly two years. These features, and the ability to tune the effective pore size of ETS-4, make it an attractive system for a range of important and difficult gas separation applications. □

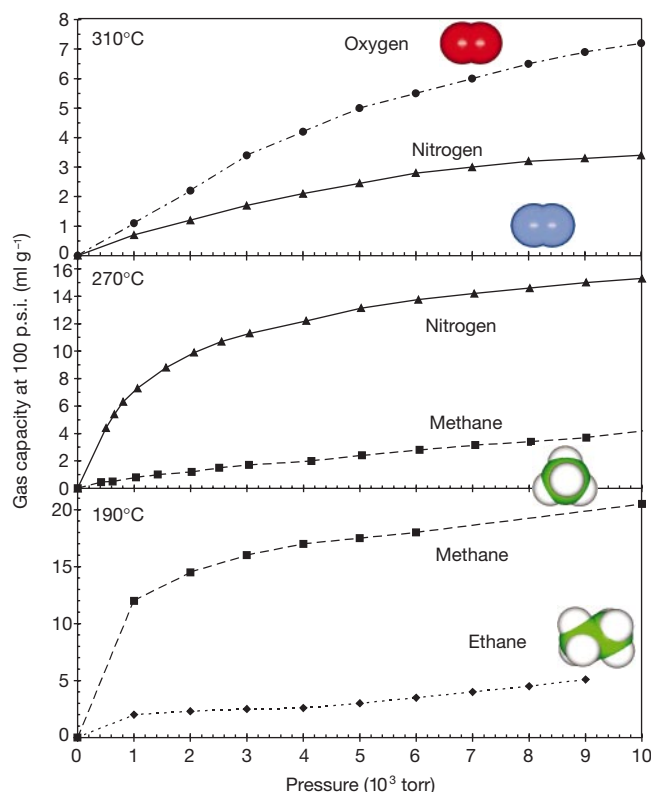


Figure 5 The molecular gate effect. Adsorption isotherms (collected at room temperature) for NaSr-ETS-4 (partially (75%) Sr-exchanged ETS-4) dehydrated at 190, 270 and 310 °C demonstrate that the corresponding CTS materials show selectivity for methane over ethane, nitrogen over methane and oxygen over nitrogen, respectively. The molecule shape illustrations were constructed using Mopac, as implemented in the Insight II (v4.0.0) software suite (Molecular Simulations Inc., San Diego, California, 1996).

Methods

Synthesis and characterization

ETS-4 synthesis and ion exchange were performed as described elsewhere^{17,21}. *In situ* X-ray diffraction was performed using a Philips X'Pert diffractometer equipped with a high-temperature stage employing a Si internal standard. TGA was performed in a DSC/TGA (differential scanning calorimetry/thermogravimetric analysis) system with a heating rate of 0.5 °C min⁻¹ in air and a flow rate of 10 litres per hour. ICP-MS (inductively coupled plasma-mass spectrometry) was used to determine the degree of ion-exchange. Adsorption data (Fig. 4) were collected in a VTI HPA-300 at 100 p.s.i. and 30 °C after dehydration for 24 h at the required temperature. Adsorption isotherms (Fig. 5) were collected, by VTI Corp., in a VTI HPA-300 at 25 °C, after 24 h of dehydration at the required temperature.

For the neutron diffraction studies, samples of Sr-exchanged ETS-4 were heat-treated at the required temperature for 24 h in Pyrex vials under dry helium flow. The vials were then sealed, isolated from the helium stream, and stored in a dry helium atmosphere. The samples were subsequently transferred to indium-sealed vanadium sample holders under a dry helium atmosphere. Neutron diffraction patterns were collected between 2θ values of 6° and 130° on the high-resolution BT-1 powder diffractometer at NIST (Gaithersburg, MD), operating at a wavelength of 2.0783 Å with a Ge(311) monochromator and a 15' Soller collimator for high neutron flux.

Structure analysis

The room-temperature structure of Sr-ETS-4 in the *Cmmm* space group (solved from X-ray data) was taken as a starting point for the solution of the contracted structure. Analysis of neutron diffraction data from the material heat treated at 150, 200, 250 and 300 °C shows that *Cmmm* symmetry is retained to a good approximation. First, structure factors were extracted from the neutron diffraction data using the Le Bail algorithm²⁷. The lattice constants (in the *Cmmm* space group), diffractometer zero, scale factor, background, and profile parameters were refined using this extraction technique before any atomic parameters were varied. The lattice constants and weighted residuals using the Le Bail fit are given in the legend of Fig. 2 and represent the best possible fit using the *Cmmm* space group.

A detailed structural analysis was performed on the samples dehydrated at 150, 200, 250 and 300 °C. Fourier difference maps were used to locate the Sr²⁺ cations in the framework for these cases. Three cation positions are located and refined in these structures. Three-site occupancy constraints are used to satisfy both the overall charge balance as well as the

total number of cations as known from the room-temperature structure. Two of the cation sites have relatively low occupancies. One of them is the original Sr²⁺ cation site in the room-temperature structure, and the other is a new site located at the centre of the 8MR, in which case the cation coordinates to the oxygen atoms of the 8MR. The third site is heavily occupied, and is also a new cation position. There is another significant peak observed in the Fourier maps which, owing to its coordination environment, is attributed to residual water still adsorbed in the framework. The oxygen (Ow) occupancy on this site decreases significantly with increasing dehydration temperature, in agreement with thermogravimetric data.

Structure refinement

After identifying these structural changes, we conducted Rietveld refinement^{28,29} of the data for the samples dehydrated at 150, 200, 250 and 300 °C. However, owing to the low number of observed diffraction peaks and disorder in the material, soft constraints were necessary throughout the refinement—elimination of these constraints results in divergences in the model. The weight of the bond-angle constraints was set to 10% of the weight of the bond-distance constraints, and results in physically realistic bond distances and bond angles. Tables 1a, 1b, 1c and 1d in the Supplementary Information show the refined cation and framework structural parameters. The lattice constants obtained from the Le Bail fit are further refined, with only minor changes occurring in their values. Figures 1a, 1b, 1c and 1d in Supplementary Information show the fitted neutron diffraction patterns for the cases considered here. The fits are reasonably good, and are close to the best possible values in the *Cmmm* space group, as obtained by the Le Bail profile fit.

Received 24 November 2000; accepted 21 June 2001.

- Hammonds, K. D., Heine, V. & Dove, M. T. Rigid-unit modes and the quantitative determination of the flexibility possessed by zeolite frameworks. *J. Phys. Chem. B* **102**, 1759–1767 (1998).
- Mentzen, B. F. & Gelin, P. The silicalite p-xylene system. 1. Flexibility of the MFI framework and sorption mechanism observed during p-xylene pore-filling by X-ray powder diffraction at room temperature. *Mater. Res. Bull.* **30**, 373–380 (1995).
- Nenoff, T. M. *et al.* Flexibility of the zeolite RHO framework. In situ X-ray and neutron powder structural characterization of cation exchanged BePO and BeAsO RHO analogs. *J. Phys. Chem.* **102**, 14256–14264 (1996).
- Johnson, G. M. *et al.* Flexibility and cation distribution upon lithium exchange of aluminosilicate and aluminogermanate materials with RHO topologies. *Chem. Mater.* **11**, 2780–2787 (1999).
- Kuznicki, S. M. Large-pored crystalline titanium molecular sieve zeolites. US Patent No. 4853202 (1989).
- Kuznicki, S. M. & Thrush, K. A. Large-pored molecular sieves with charged octahedral titanium and charged tetrahedral aluminum sites. US Patent No. 5244650 (1993).
- Kuznicki, S. M. & Thrush, K. A. Large-pored molecular sieves containing at least one octahedral site comprising titanium and at least silicon as a tetrahedral site. US Patent No. 5208006 (1993).
- Kuznicki, S. M. Preparation of small-pored crystalline titanium molecular sieve zeolites. US Patent No. 4938939 (1990).
- Anderson, M. W. *et al.* Structure of the microporous titanosilicate ETS-10. *Nature* **367**, 347–351 (1994).
- Anderson, M. W. *et al.* Microporous titanosilicate ETS-10, a structural survey. *Phil. Mag. B* **71**, 813–841 (1995).
- Wang, X. & Jacobson, A. J. Crystal structure of the microporous titanosilicate ETS-10 refined from single crystal X-ray diffraction data. *Chem. Commun.* 973–974 (1999).
- Anderson, M. W. *et al.* Isomorphous substitution in the microporous titanosilicate ETS-10. *Microporous Mater.* **6**, 195–204 (1996).
- Eldewik, A., Luca, V., Singh, N. K. & Howe, R. F. Iron substitution in the microporous titanosilicate ETS-10. *Proc. Int. Zeolite Conf.* **3**, 1507–1514 (1998).
- Lamberti, C. Electron-hole reduced effective mass in monoatomic -OTiOTiO- quantum wires embedded in the siliceous crystalline matrix of ETS-10. *Microporous Mesoporous Mater.* **30**, 155–163 (1999).
- Rocha, J. & Anderson, M. W. Microporous titanosilicates and other novel mixed octahedral-tetrahedral framework oxides. *Eur. J. Inorg. Chem.* 801–818 (2000).
- Bordiga, S. *et al.* Stoichiometric and sodium-doped titanium silicate molecular sieve containing atomically defined -OTiOTiO- chains: Quantum ab initio calculations, spectroscopic properties, and reactivity. *J. Chem. Phys.* **112**, 3859–3867 (2000).
- Kuznicki, S. M., Bell, V. A., Petrovic, I. & Desai, B. T. Small-pored crystalline titanium molecular sieve zeolites and their use in gas separation processes. US Patent No. 6068682 (2000).
- Sandomirskii, P. A. & Belov, N. V. The OD structure of zorite. *Sov. Phys. Crystallogr.* **24**, 686–693 (1979).
- Mer'kov, A. N. *et al.* Raite and zorite, new minerals from the Lovozero tundra. *Zap. Vses. Mineral. Obshchest.* **102**, 54–62 (1973).
- Philippou, A. & Anderson, M. W. Structural investigation of ETS-4. *Zeolites* **16**, 98–107 (1996).
- Braunbarth, C. *et al.* Structure of strontium ion exchanged ETS-4 microporous molecular sieves. *Chem. Mater.* **12**, 1857–1865 (2000).
- McCusker, L. B. in *Comprehensive Supramolecular Chemistry, Solid-State Supramolecular Chemistry: Two- and Three-Dimensional Inorganic Networks* (eds Alberti, G. & Bein, T.) 393–423 (Pergamon, New York, 1996).
- Cruciani, G., De Luca, P., Nastro, A. & Pattison, P. Rietveld refinement of the zorite structure of ETS-4 molecular sieves. *Microporous Mesoporous Mater.* **21**, 143–153 (1998).
- Bieniok, A. & Hammonds, K. D. Rigid unit modes and the phase transition and structural distortions of zeolite rho. *Microporous Mesoporous Mater.* **25**, 193–200 (1998).
- Bieniok, A. & Baur, W. H. A large volume contraction accompanies the low- to high-temperature transition of zeolite Sr-RHO. *J. Solid State Chem.* **90**, 173–182 (1991).
- Depmeier, W. in *Molecular Sieves Vol. 2, Structures and Structure Determination* (eds Karge, H. J. & Weitkamp, J.) 113–140 (Springer, Berlin, 1999).
- Le Bail, A., Duroy, H. & Fourquet, J. L. Ab-initio structure determination of LiSiWO₆ by X-ray powder diffraction. *Mater. Res. Bull.* **23**, 447–452 (1988).

28. Young, R. A. (ed.) *The Rietveld Method* (Oxford Univ. Press, Oxford, 1993).
29. Larson, A. C. & von Dreele, R. B. *GSAS Report LAUR 86-784* (Los Alamos National Laboratory, 1986).
30. Breck, D. W. *Zeolite Molecular Sieves* (Wiley, New York, 1974).
31. Mitriten, M. & Dolan, W. Nitrogen removal from natural gas with molecular gate technology. *Proc. Laurence Reid Gas Cond. Conf.* **51**, 1–16 (2001).
32. Abrams, L. & Corbin, D. R. in *Topics in Inclusion Science Vol. 6, Inclusion Chemistry with Zeolites: Nanoscale Materials by Design* (eds Herron, N. & Corbin, D. R.) 4–6 (Kluwer, Dordrecht, 1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. Curran for discussions and assistance in the preparation of the manuscript, and VTI Corp. for the data of Fig. 5. This work was supported by ATP/NIST, the David and Lucile Packard Foundation and NSF-CTS.

Correspondence and requests for materials should be addressed to S.M.K. (e-mail: steve.kuznicki@engelhard.com).

The timing of the last deglaciation in North Atlantic climate records

Claire Waelbroeck*, Jean-Claude Duplessy*, Elisabeth Michel*, Laurent Labeyrie†, Didier Paillard* & Josette Duprat‡

* Laboratoire des Sciences du Climat et de l'Environnement, Domaine du CNRS, bât. 12, 91198 Gif-sur-Yvette, France

† Département des Sciences de la Terre, Université Paris-sud Orsay, bât. 504, 91104 Orsay, France

‡ Département de Géologie et Océanographie, CNRS UMR 5805, Université de Bordeaux I, Talence, France

To determine the mechanisms governing the last deglaciation and the sequence of events that lead to deglaciation, it is important to obtain a temporal framework that applies to both continental and marine climate records. Radiocarbon dating has been widely used to derive calendar dates for marine sediments, but it rests on the assumption that the 'apparent age' of surface water (the age of surface water relative to the atmosphere) has remained constant over time^{1,2}. Here we present new evidence for variation in the apparent age of surface water (or reservoir age) in the North Atlantic ocean north of 40°N over the past 20,000 years. In two cores we found apparent surface-water ages to be larger than those of today by $1,230 \pm 600$ and $1,940 \pm 750$ years at the end of the Heinrich 1 surge event (15,000 years BP) and by 820 ± 430 to $1,010 \pm 340$ years at the end of the Younger Dryas cold episode. During the warm Bølling–Allerød period, between these two periods of large reservoir ages, apparent surface-water ages were comparable to present values. Our results allow us to reconcile the chronologies from ice cores and the North Atlantic marine records over the entire deglaciation period. Moreover, the data imply that marine carbon dates from the North Atlantic north of 40°N will need to be corrected for these highly variable effects.

The last deglaciation is a period particularly well dated because it belongs to the time span covered by isotopic (¹⁴C) dating techniques, which can be applied to both continental and marine sediments. Moreover, ¹⁴C dates can be converted into calendar ages through calibration curves describing the variations in the atmospheric ¹⁴C/¹²C ratio over the past 20 kyr (ref. 3). Also, the uncertainty on the absolute dating by annual-layer counting of the Greenland ice cores is small (<550 yr) over the last 15 kyr (ref. 4). Therefore, it is possible to compare directly the timing of events from continental and ice-core records, and thus to infer causal mechanisms from the data. The situation is more complex for the marine data, as the ¹⁴C dates are measured on foraminifera preserved in the sediments and hence reflect the ¹⁴C/¹²C ratio of the

water in which the foraminifera calcified. The surface-water ¹⁴C/¹²C ratio is different from that of the contemporaneous atmosphere, reflecting the balance between the input of atmospheric ¹⁴C and its removal by transport and radiodecay in the water column. This difference in ¹⁴C/¹²C ratio is usually expressed as the apparent or reservoir age of the water mass.

A compilation of pre-bomb surface-water ¹⁴C content indicates that the modern surface reservoir age is about 400 ± 100 yr in the tropics and in the North Atlantic whereas it rises to 1,200 yr at higher latitudes in the Southern and North Pacific oceans^{5,6}. Past reservoir ages most probably differed from those of today, but only sparse data exist. Reservoir ages have been measured in a few sites at given times in the past, by dating contemporaneous samples in marine sediments and in terrestrial organic matter, marked by the same volcanic ash layer^{7–10}.

Here, we present summer sea surface temperature (SST) reconstructions and benthic oxygen isotopic records ($\delta^{18}\text{O}_b$) from three sediment cores raised for the North Atlantic Ocean between 37 and 55°N (SU 81-18, 37°46'N, 10°11'W, 3,135 m; CH 69-09, 41°45'N, 47°21'W, 4,100 m; and NA 87-22, 55°29'N, 14°41'W, 2,161 m). These records have been dated by accelerator mass spectrometry on monospecific planktonic foraminifera samples^{1,11,12} (see Supplementary Information Table 1). Radiocarbon ages have been converted into calendar ages with the CALIB 4.1 software¹³, the smoothed (310-yr moving average) 1998 marine calibration curve

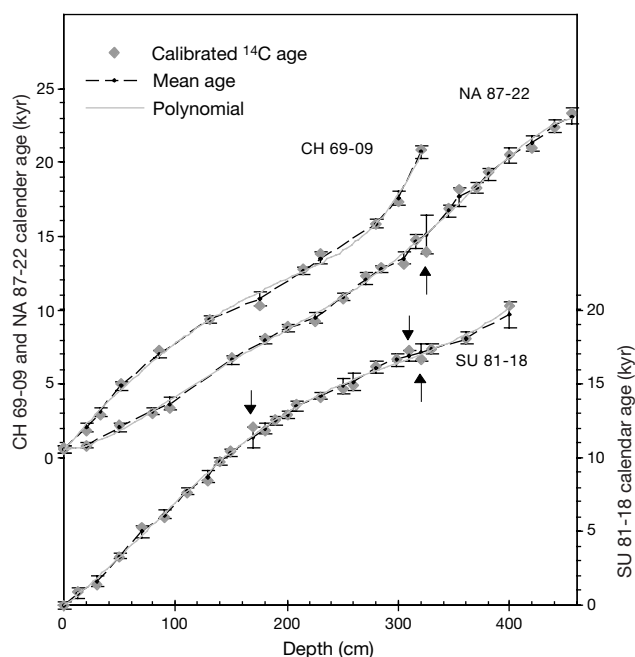


Figure 1 Age–depth relationships for the three North Atlantic deep sea cores. Calendar ages were calculated from the measured ¹⁴C ages (see Supplementary Information Table 1) with the CALIB 4.1 software¹³ and a constant surface reservoir age of 400 yr. For each core, two different interpolation schemes were used: first, we computed the fifth-order polynomial fitting the dated levels (grey line); second, we linearly interpolated the dated levels, leaving out one age in core NA 87-22 and three ages in core SU 81-18 that showed small age inversions (arrows). The error on the linearly interpolated age scale is taken as one standard deviation of the sample's calendar-age probability distribution given by CALIB 4.1. The error on the polynomial age scale is taken as the difference between the actual sample calendar age and the polynomial age computed at the same level, when this difference is larger than the error on the sample's calendar age. It is otherwise equal to the error on the linearly interpolated age scale. We derived the final age model for each core (black dots with 1-s.d. error bars) by taking the arithmetical mean of the above two timescales to ensure realistically large error bars. The final error estimate includes an additional 200-yr uncertainty, representing the maximum possible bioturbation bias. Both interpolation schemes support our conclusions.

of Stuiver *et al.*³, and a constant surface reservoir age of 400 yr (Fig. 1). The method provides the probability distribution of the sample's calendar age.

The sedimentation rate is about 20 cm kyr⁻¹ between 10 and 15 kyr before present (BP) for the three cores examined here (Fig. 1). Therefore, bioturbation should not induce significant biases¹⁴. Moreover, the ¹⁴C ages were determined on planktonic monospecific samples selected in the species abundance maxima (see Supplementary Information Table 1), so that bioturbation distortion of the above dating is minimized.

The three North Atlantic SST series can now be plotted against calendar age, together with the GRIP or GISP2 ice–oxygen isotopic record, a proxy for air temperature¹⁵ (Fig. 2). All three SST series depict a clear two-step warming of 7–12 °C amplitude from the end of the last glacial to the Holocene. The first warming phase is extremely abrupt, as can be seen in the two records of highest resolution (CH 69-09 and SU 81-18) and corresponds to GRIP's sharp increase in air temperature marking the transition between the last Heinrich event (H1) (ref. 16) and the Bølling–Allerød interval. Similarly, the second rapid-warming phase should correspond with the transition from the Younger Dryas cold episode to the Holocene.

Assuming a constant surface reservoir age implies very large leads of the SST increase at the two northernmost Atlantic sites with

respect to the increase in air temperature above Greenland. These leads can be accurately evaluated (see Supplementary Information Table 2), with the total uncertainty on the warming-midpoint dates resulting from the error related to the time resolution and from the dating error (Fig. 1). The first sharp SST increase is found to lead the H1–Bølling–Allerød transition recorded in the Greenland ice by 1.94 ± 0.75 kyr in core NA 87-22, and by 1.23 ± 0.60 kyr in core CH 69-09. Similarly, the second abrupt SST increase leads the Younger Dryas–Holocene transition in the ice by 0.82 ± 0.43 kyr in core NA 87-22, and 1.01 ± 0.34 kyr in core CH 69-09. In contrast, the SST variations recorded in core SU 81-18, located further south (38° N), coincide with air temperature variations above Greenland.

Several lines of evidence drive us to conclude that the large leads observed in the northernmost records are not real climatic features. First, examining the marine $\delta^{18}\text{O}_b$ records reveals a similar significant lead of the decrease in $\delta^{18}\text{O}_b$ associated with the deglaciation in the northern records compared with the SU 81-18 mid-latitudes $\delta^{18}\text{O}_b$ record (Fig. 3a). The decrease in $\delta^{18}\text{O}_b$ taking place during deglaciations results from the combined effect of ice melting and deep-water warming. Because the three cores are geographically relatively close and bathed by water masses dominated by North Atlantic deep water (NADW), large changes in $\delta^{18}\text{O}_b$ should take place roughly simultaneously at the three locations. Therefore, a different timing of the warming cannot explain the lead of 1–2 kyr

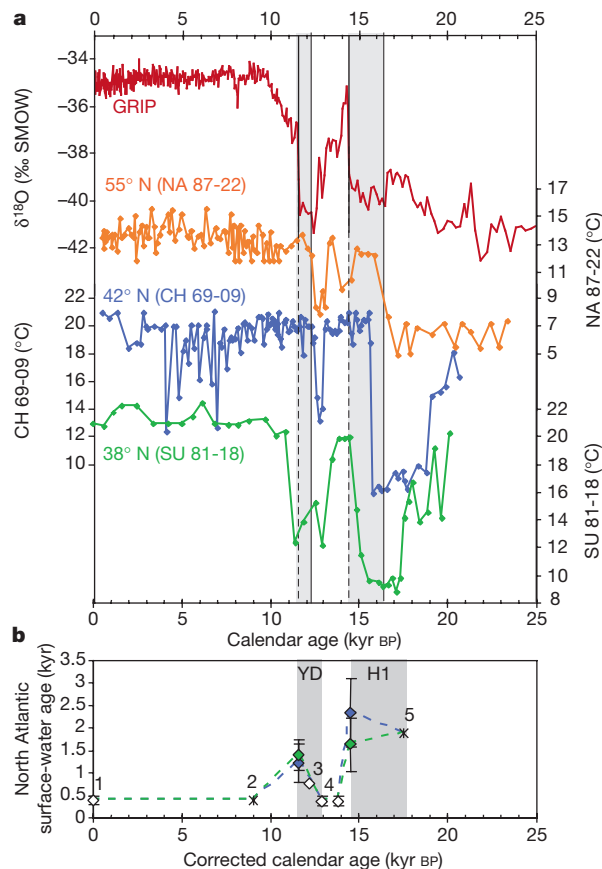


Figure 2 Surface temperature records and reconstructed evolution of the surface reservoir age during the last 18 kyr. **a**, Comparison of GRIP ice $\delta^{18}\text{O}$ (a proxy for air temperature) and summer SST reconstructions at the three North Atlantic sites versus calendar age. SST has been reconstructed from the planktonic foraminifera abundances by the revised analogue method^{27,28}. Grey stripes highlight the large leads between NA 87-22 SST and GRIP air temperature at the H1–Bølling–Allerød and Younger Dryas–Holocene transitions. **b**, Tentative synthesis of variations in surface-water ages in the North Atlantic north of 40° N during the last 18 kyr. Filled diamonds, our results; unfilled

diamonds, estimates available in the literature: 1, pre-bomb value^{5,6}; 3, reservoir ages obtained by dating of the Vedde ash layer^{7,8}; 4, Bølling–Allerød reservoir ages measured in Norwegian coastal waters²⁹. Crosses, possible additional constraints: 2, ocean circulation in the North Atlantic can be considered as similar to today around 9 kyr BP (ref. 20); 5, approximate value derived from the difference between the initiation of the cooling associated with H1 in SU 81-18 and in CH 69-09, assuming a reservoir age of 0.4 kyr at the SU 81-18 site. Dotted line, proposed evolution of surface-water age at high latitudes in the North Atlantic.

of the northern $\delta^{18}\text{O}_b$ records with respect to the $\delta^{18}\text{O}_b$ signal of core SU 81-18. The lead of the northern records with respect to the mid-latitudes record seems thus to be an artefact of the ^{14}C chronology.

Second, no reasonable climatic mechanism can explain how SST at site NA 87-22, located in the same latitude band as Greenland, could have reached interglacial values 0.8–1.9 kyr before the SST off Portugal (site SU 81-18) and the air temperature above Greenland.

On the contrary, we have reasons to believe that the small surface-water ages currently observed in the North Atlantic did not prevail in the past. In fact, current constant reservoir ages over the North Atlantic are principally due to the North Atlantic Drift, which brings low-latitude surface waters to 50–60°N and into the Norwegian Sea. The cooling of warm and saline surface waters at high latitudes induces active deep-water formation in these areas, associated with a large flux of atmospheric ^{14}C to the surface waters. Consequently, the presence of large continental ice caps, the southward shift of the polar front and the associated shift in deep-water-formation areas during the Younger Dryas cold episode and the last glacial maximum^{17–21} probably explain an important decrease in ventilation and hence an increase in surface-water reservoir age, north of the polar front⁷. It was also demonstrated that iceberg discharge H1 had a large impact on ocean circulation, further reducing NADW formation^{20,22}.

The effect of circulation changes on the ocean radiocarbon content has been evaluated with coupled ocean–ice–atmosphere models^{23–25}. These modelling experiments show that a reduction in NADW overturn and an increase in sea ice cover lead to an increase in surface-water age of 200–300 yr in the Atlantic, north of 40°N. These results are thus consistent with our findings, although the increase in surface reservoir age predicted by the models is weaker than our data indicate. This seems to result from the fact that these models simulate no or very little vertical convection north of the polar front, although palaeoceanographic data show the opposite²⁶.

The two northernmost cores demonstrated leads of the SST abrupt warmings with respect to the air temperature increases recorded in the Greenland ice. We interpret these leads as increases in the surface reservoir age at the cores' location during these two specific climatic transitions. A tentative reconstruction of the

evolution of the Atlantic surface-water age during the last 18 kyr can be derived from the compilation of available estimates of apparent surface-water ages (Fig. 2b). After correcting the marine time scales to account for these large reservoir ages, the three $\delta^{18}\text{O}_b$ records are in much better agreement between 11 and 18 kyr (Fig. 3b), the residual differences probably resulting from measurement uncertainties, noise due to bioturbation, and small differences in deep water temperature variations between sites.

This study allowed us to reconcile ice core and North Atlantic marine chronologies, as well as surface and deep-water climatic records from mid- to high latitudes, over the entire deglaciation period. We demonstrated that the surface-water age at high latitudes in the North Atlantic has varied by a factor of 3–5 during the past 15 kyr, owing presumably to major changes in ocean circulation and sea ice cover. It is therefore highly probable that it also significantly varied across the Heinrich and Dansgaard–Oeschger climatic events of the last glacial period (see for example ref. 2), inducing biases in marine ^{14}C chronologies. The situation at high southern latitudes is expected to be similar, as indicated by recent measurements showing large increases in surface reservoir ages in the New Zealand region during glacial times¹⁰. □

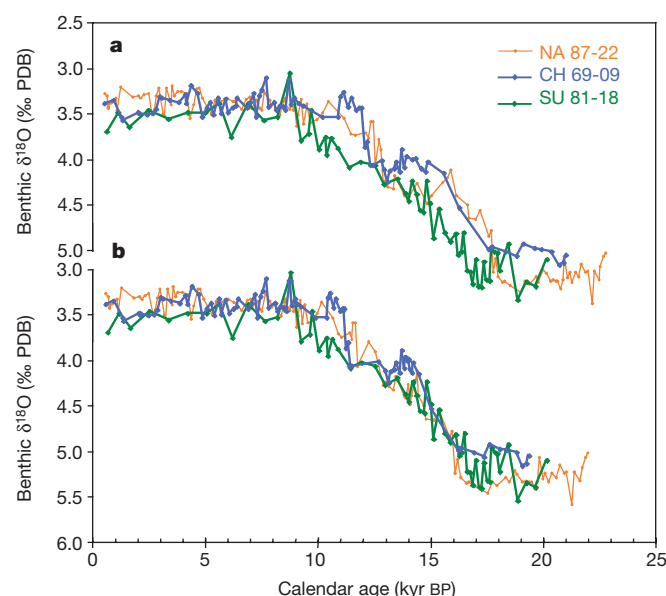


Figure 3 $\delta^{18}\text{O}_b$ records from the three North Atlantic cores. Data^{11,12,30} are expressed in ‰ versus the Pee Dee Belemnite (PDB) standard. **a**, Records versus the calendar-age scale defined in Fig. 1. **b**, Same records after correcting the original age scale to account for the variations in surface-water age according to our tentative synthesis (Fig. 2b).

- Bard, E. *et al.* Retreat velocity of the North Atlantic polar front during the last deglaciation determined by ^{14}C accelerator mass spectrometry. *Nature* **328**, 791–794 (1987).
- Voelker, A. *et al.* Correlation of marine ^{14}C ages from the Nordic Seas with the GISP2 isotope record: Implications for ^{14}C calibration beyond 25 ka BP. *Radiocarbon* **40**, 517–534 (1998).
- Stuiver, M. *et al.* INTCAL98 radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* **40**, 1041–1083 (1998).
- Alley, R. B. *et al.* Abrupt increase in Greenland snow accumulation at the end of the Younger Dryas event. *Nature* **362**, 527–529 (1993).
- Bard, E. Correction of accelerator mass spectrometry ^{14}C ages measured in planktonic foraminifera: paleoceanographic implications. *Paleoceanography* **3**, 635–645 (1988).
- Stuiver, M. & Braziunas, T. F. Sun, ocean, climate and atmospheric $^{14}\text{CO}_2$: an evaluation of causal and spectral relationship. *The Holocene* **3**, 289–305 (1993).
- Bard, E. *et al.* The North Atlantic atmosphere–sea surface ^{14}C gradient during the Younger Dryas climatic event. *Earth Planet. Sci. Lett.* **126**, 275–287 (1994).
- Austin, W. E. N., Bard, E., Hunt, J. B., Kroon, D. & Peacock, J. D. The ^{14}C age of the Icelandic Vedde Ash: implications for Younger Dryas marine reservoir age corrections. *Radiocarbon* **37**, 53–62 (1995).
- Siani, G. *Estimation de l'âge ^{14}C du réservoir des eaux de surface de la mer Méditerranée pendant les derniers 18000 ans*. Thesis, Univ. Paris-Sud (1999).
- Sikes, E. L., Samson, C. R., Guilderson, T. P. & Howard, W. R. Old radiocarbon ages in the southwest Pacific Ocean during the last glacial period and deglaciation. *Nature* **405**, 555–559 (2000).
- Duplessy, J. C. *et al.* Changes in surface salinity of the North Atlantic Ocean during the last deglaciation. *Nature* **358**, 485–487 (1992).
- Labeyrie, L. *et al.* in *Mechanisms of Global Climate Change at Millennial Time Scales*, *Geophysical Monograph Series* 112 (eds Clark, P. & Webb, R. S.) 77–98 (American Geophysical Union, Washington, DC, 1999).
- Stuiver, M. & Reimer, P. J. Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon* **35**, 215–230 (1993).
- Bard, E., Arnold, M., Duprat, J., Moyes, J. & Duplessy, J. C. Reconstruction of the last deglaciation: deconvolved records of $\delta^{18}\text{O}$ profiles, micropaleontological variations and accelerator mass spectrometric ^{14}C dating. *Clim. Dynam.* **1**, 101–112 (1987).
- Johnsen, S. J. *et al.* The $\delta^{18}\text{O}$ record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climatic instability. *J. Geophys. Res.* **102**, 26397–26410 (1997).
- Broecker, W., Bond, G., Klas, M., Clark, E. & MacManus, J. Origin of the northern Atlantic's Heinrich events. *Clim. Dynam.* **6**, 265–273 (1992).
- Boyle, E. A. Cadmium: chemical tracer deepwater paleoceanography. *Paleoceanography* **3**, 471–489 (1988).
- Labeyrie, L. D. *et al.* Changes in vertical structure of the North Atlantic Ocean between glacial and modern times. *Quat. Sci. Rev.* **11**, 401–413 (1992).
- Duplessy, J. C. *et al.* Deepwater source variations during the last climatic cycle and their impact on the global deepwater circulation. *Paleoceanography* **3**, 343–360 (1988).
- Sarnthein, M. *et al.* Changes in east Atlantic deepwater circulation over the last 30,000 years: eight time slice reconstructions. *Paleoceanography* **9**, 209–267 (1994).
- Muscheler, R., Beer, J., Wagner, G. & Finkel, R. Changes in deep-water formation during the Younger Dryas event inferred from ^{10}Be and ^{14}C records. *Nature* **408**, 567–570 (2000).
- Vidal, L. *et al.* Evidence for changes in the North Atlantic Deep Water linked to meltwater surges during the Heinrich events. *Earth Planet. Sci. Lett.* **146**, 13–26 (1997).
- Mikolajewicz, U. A meltwater induced collapse of the conveyor belt thermohaline circulation and its influence on the distribution of $\Delta^{14}\text{C}$ and $\delta^{18}\text{O}$ in the oceans. *Max Planck Inst. Rep.* **189**, 1–25 (1996).
- Stocker, T. F. & Wright, D. G. Rapid changes in ocean circulation and atmospheric radiocarbon. *Paleoceanography* **11**, 773–795 (1996).
- Stocker, T. F. & Wright, D. G. The effect of a succession of ocean ventilation changes on ^{14}C . *Radiocarbon* **40**, 359–366 (1998).
- Dokken, T. M. & Jansen, E. Rapid changes in the mechanism of ocean convection during the last glacial period. *Nature* **401**, 458–461 (1999).
- Waelbroeck, C. *et al.* Improving past sea surface temperature estimates based on planktonic fossil faunas. *Paleoceanography* **13**, 272–283 (1998).

28. Malmgren, B., Kucera, M., Nyberg, J. & Waelbroeck, C. Comparison of available statistical and artificial neural network techniques for estimating past sea-surface temperatures from planktonic foraminifera census data. *Palaeoceanography* (in the press).
29. Bondevik, S., Birks, H. H., Gulliksen, S. & Mangerud, J. Late Weichselian marine ^{14}C reservoir ages at the western coast of Norway. *Quat. Res.* **52**, 104–114 (1999).
30. Labeyrie, L. *et al.* Surface and deep hydrology of the northern Atlantic Ocean during the last 150,000 years. *Phil. Trans. R. Soc. Lond.* **348**, 255–264 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank E. Bard and G. Siani for discussions and comments. The data were acquired in collaboration with T. van Weering, J.-L. Turon, M. Labracherie and G. Auffret. We thank B. Lecoat and J. Tessier for processing the isotopic measurements. This study was supported by CNRS, CEA, INSU (PNEDC) and EU Environment Program.

Correspondence and requests for materials should be addressed to C.W. (e-mail: Claire.Waelbroeck@lsce.cnrs-gif.fr).

Direct observation of a submarine volcanic eruption from a sea-floor instrument caught in a lava flow

Christopher G. Fox^{*†}, William W. Chadwick Jr^{†‡} & Robert W. Embley^{*}

^{*} NOAA/PMEL, Newport, Oregon 97365, USA

[‡] 2115 SE OSU Drive, Oregon State University/NOAA, Newport, Oregon 97365, USA

[†] These authors contributed equally to this work

Our understanding of submarine volcanic eruptions has improved substantially in the past decade owing to the recent ability to remotely detect such events¹ and to then respond rapidly with synoptic surveys and sampling at the eruption site. But these data are necessarily limited to observations after the event². In contrast, the 1998 eruption of Axial volcano on the Juan de Fuca ridge^{3,4} was monitored by *in situ* sea-floor instruments^{5–7}. One of these instruments, which measured bottom pressure as a proxy for vertical deformation of the sea floor, was overrun and entrapped by the 1998 lava flow. The instrument survived—being insulated from the molten lava by the solidified crust—and was later recovered. The data serendipitously recorded by this instrument reveal the duration, character and effusion rate of a sheet flow eruption on a mid-ocean ridge, and document over three metres of lava-flow inflation and subsequent drain-back. After the brief two-hour eruption, the instrument also measured gradual subsidence of 1.4 metres over the next several days, reflecting deflation of the entire volcano summit as magma moved into the adjacent rift zone. These findings are consistent with our understanding of submarine lava effusion, as previously inferred from seafloor observations, terrestrial analogues, and laboratory simulations^{8–11}.

Two Volcanic System Monitor (VSM) instruments were deployed at Axial volcano in October 1997, one near the centre of the caldera⁵ and one on the upper south rift zone, hereafter referred to as VSM1 and VSM2, respectively (Figs 1 and 2). The VSM instruments measure ambient pressure on the sea floor every 15 s with a Paroscientific digiquartz pressure sensor. The data processing methods are described in detail elsewhere^{5,12}, but include filtering to remove oceanographic phenomena and a correction for temperature. In August 1998, VSM2 was found to be trapped in the new lava flow (Fig. 1), but it was pulled free the following summer.

The rescued instrument was in surprisingly good condition. The

maximum temperature recorded inside VSM2 during the eruption was only 7.5 °C (Fig. 3)—remarkably low, considering that the instrument was sitting atop basaltic lava that was probably erupted at ~1,190 °C (M. Perfit, personal communication). This was apparently due to the thermal insulation provided by the surface crust that forms (and quickly thickens) when submarine lava flows come into contact with frigid sea water¹³.

The northern end of the 1998 lava flow where VSM2 was located (Fig. 2) is a long, narrow sheet flow⁴. Submarine sheet flows are thought to form by brief, relatively high-effusion-rate eruptions^{8,9}. Recent models^{10,11} suggest that such flows initially advance as thin lobate flows (20–30 cm thick), spreading at a rate that is rapid enough for the interior of the flow to remain a single fluid core beneath a solid upper crust. The flow spreads until it becomes laterally confined, and then 'inflates' upward. During inflation, the upper crust is uplifted (sometimes > 5 m) by hydraulic pressure within the flow interior—and lava pillars grow upward, connecting the upper and lower crusts¹⁰. Lava inflation continues until the eruption rate wanes or stops, at which point lava begins to drain back and the upper crust founders where it is left unsupported, leaving extensive areas of collapse within the flow.

Geologic mapping of the 1998 lava flow shows that VSM2 was located in the collapsed interior of the 1998 lava flow, ~3 m below nearby remnants of the uncollapsed upper crust and ~160 m west of the eruptive fissure (Fig. 2). The instrument had not been buried by lava, even though the lava level had been 3 m higher before collapse. This observation implies that lava initially flowed under the instrument, which became embedded in the upper crust of the flow when the flow subsequently inflated. During inflation, VSM2 was uplifted along with the upper crust (because of its table-like design) and then was set back down within the collapse area during the drain-back stage.

The dataset recovered from the VSM2 instrument directly recorded this sequence of events (Fig. 3). Uplift of the instrument began slowly at 14:55 on 25 January 1998 (all times GMT), after the onset of the earthquake swarm at 11:33 (ref. 3) and subsidence began at VSM1 at 14:51 (ref. 5). Rapid inflation abruptly started at 15:16 when VSM2 was uplifted 168 cm in the next 5 minutes, an average rate of 32 cm min⁻¹ (although instantaneous rates were as high as 73 cm min⁻¹). After a pause of 4 minutes, during which the instrument subsided 13 cm, inflation resumed for another 21 minutes which uplifted the instrument another 113 cm at a decreased rate of 5.5 cm min⁻¹. At 15:46, inflation dramatically slowed, probably owing to a sudden reduction in the effusion rate at the vent, and the instrument was only uplifted another 7 cm over the next 21 minutes. During this period the lava flow remained at or near its fully inflated thickness, and the recorded temperature reached its peak of 7.5 °C at 15:49 (Fig. 3).

At this point, lava inflation ended and lava drain-back began, presumably reflecting the end of lava effusion at the vent. Our

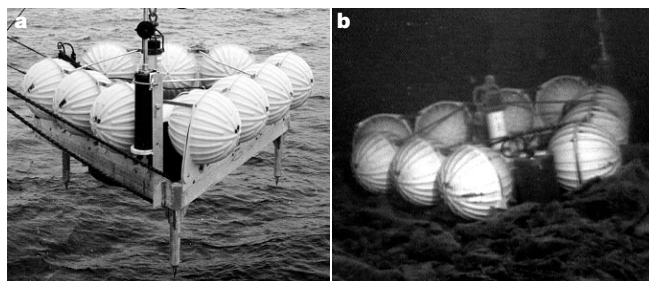


Figure 1 Views of the VSM2 instrument. **a**, At the surface during deployment (legs are 46 cm long), and **b**, on the sea floor, stuck in the 1998 lava flow at Axial volcano. Virtual animations and video clips of VSM2 (referred to as a 'rubbleometer') on the sea floor are available at <http://www.pmel.noaa.gov/vents/nemo/explorer/rubble.html>

interpretation is that the lava then drained directly back down into the vent, because the 1998 flow is relatively narrow here and because the slope direction within the collapse area is directly toward the eruptive fissure (Fig. 2). During lava drain-back, VSM2 subsided 281 cm, at a nearly uniform rate of 3.5 cm min^{-1} (Fig. 3). The recorded temperature also declined during drain-back, reaching 5.2°C at 17:16. When drain-back ended at 17:28, the instrument was only 24 cm higher than when it started, but over the next several hours it regained about 30 cm of lost elevation, perhaps during the waning stages of the eruption. A second peak in temperature (7.3°C) was recorded at 18:12, following the end of drain-back, after which the temperature steadily declined. In the end, only 1 m of lava remained in the floor of the collapse area where the flow had previously inflated to a thickness of 3.5 m. The entire inflation/deflation cycle lasted only 2 hours and 33 minutes, but the duration of active lava effusion from the eruptive vent was probably only 72 minutes (the inflation phase).

The 1998 lava flow is $\sim 350 \text{ m}$ wide in the east–west direction at the latitude of the position of VSM2 (Fig. 2). The eruptive fissure is linear and roughly oriented north–south, as are the east and west edges of the lava flow. This simple flow geometry allows us to calculate volumetric rates of both eruption and drain-back in the vicinity of the VSM2 instrument, based on the pressure data. To make these estimates we assume that the lava flow is two-dimensional, exactly 300 m wide with vertical sides, and that the VSM2 data represent the lava level within the entire width of the flow. This approximation is reasonable as the flow is actually wider than 300 m , but tapers near its edges.

All the rates discussed below are expressed as a volumetric rate per 1-m length of eruptive fissure, as the eruption was from a line-source rather than a point-source. We have no way of estimating the initial effusion rate while lava spread out from the vent as a thin sheet before VSM2 began to record inflation. Once inflation began, a volumetric effusion rate of $2.7 \text{ m}^3 \text{ s}^{-1}$ is associated with the time period of the highest rate of lava inflation. The effusion rate then

decreased by an order of magnitude to $0.27 \text{ m}^3 \text{ s}^{-1}$ during the slower second half of the inflation phase, and then dropped another order of magnitude to $0.016 \text{ m}^3 \text{ s}^{-1}$ during the interval when the flow was fully inflated. During drain-back, the rate that lava was flowing back into the vent was $0.17 \text{ m}^3 \text{ s}^{-1}$. This pattern of a brief burst at a high effusion rate followed by a longer period of waning rates is common at basaltic volcanoes¹⁴.

Laboratory analogue experiments have previously been used to associate the morphology of submarine lava flows with the effusion rates that may have produced them^{8,9}. The work we report here provides, we believe for the first time, a means to estimate directly the effusion rate for a submarine eruption—and we can compare these results with those predicted by the experimental models. The 1998 lava developed a lobate flow morphology as it spread out from the vent near VSM2. Geochemical analyses of the 1998 lava suggest that its viscosity at eruption was $\sim 30 \text{ Pa s}$ (M. Perfit, personal communication). If we assume that the effusion rate during this initial stage was the same as during the beginning of flow inflation ($\sim 3 \text{ m}^3 \text{ s}^{-1}$ per unit length of the vent), then Axial's 1998 flow would be consistent with the predictions from previous laboratory results^{8,9}, because it plots within the 'lobate' field, between the 'pillowed' and 'folded' domains (where the value of the dimensionless physical parameter, Ψ , is 3–13).

It is difficult to extrapolate these volumetric effusion rates per unit length of eruptive fissure to total eruption rates; this is because the VSM2 data only constrain the rates in the vicinity of the instrument, the eruptive fissure for the northern portion of the 1998 Axial eruption was probably $\sim 2.6 \text{ km}$ long, and the discharge rate and duration may have varied along its length. Nevertheless, if we assume the eruptive conditions were uniform along the entire eruptive fissure, the instantaneous total eruption rate would have

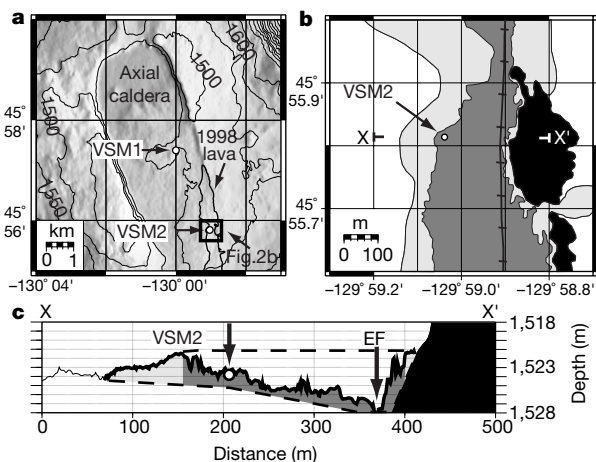


Figure 2 Location of the VSM2 instrument. **a**, Map of the caldera at Axial volcano, showing locations of the VSM1 and VSM2 instruments and the outline of the 1998 lava flow. Location of **b** is indicated. **b**, Detailed geologic map showing the location of VSM2 within the 1998 lava flow, based on scanning sonar bathymetry and bottom observations. Areas shown light grey and dark grey are respectively uncollapsed and collapsed parts of the 1998 flow; areas shown white and black are respectively older surrounding sheet flows and high-standing pillow flows. Hatched line is the 1998 eruptive fissure. X–X' is location of cross-section in **c**. **c**, Cross-section showing location of VSM2 within the collapsed interior of the 1998 lava flow, $\sim 3 \text{ m}$ below remnants of the flow's upper crust, and yet unburied by lava. Depth is shown as metres below the sea surface. EF, 1998 eruptive fissure. Dashed lines show probable upper and lower extent of 1998 lava flow before collapse. Shading as in **b**.

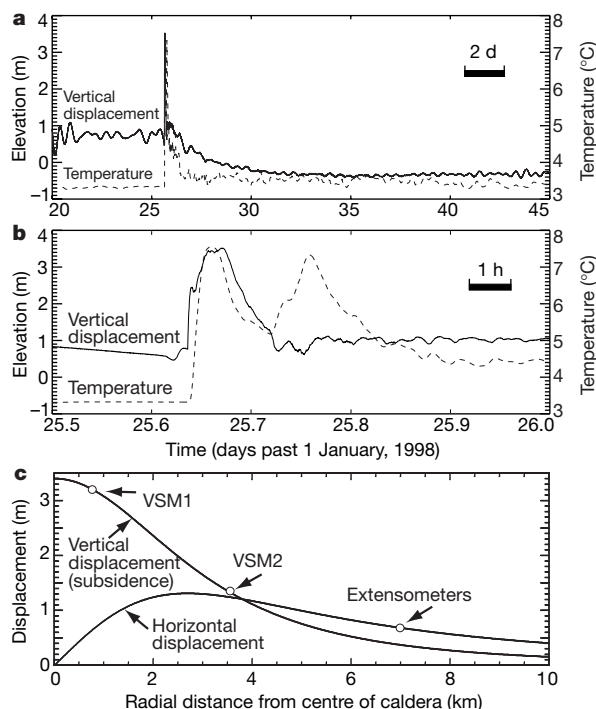


Figure 3 Data from the VSM2 instrument. **a**, Vertical displacement (solid line; from pressure data) and temperature (dashed line; measured inside a pressure case) from 20 January to 14 February 1998. **b**, Details of the vertical displacement and temperature records at the time of the eruption on 25 January. Zero elevation in **a** and **b** is arbitrary. **c**, Elastic point source deformation model¹⁹ showing how the VSM2 data fitted the curves previously derived from VSM1 and extensometer instruments⁶. Model assumes point-source is located 3.8 km beneath the centre of the caldera.

been as high as $7,000 \text{ m}^3 \text{ s}^{-1}$ early in the eruption, or $\sim 550 \text{ m}^3 \text{ s}^{-1}$ if averaged over the 72 minutes of flow inflation. These numbers are at the upper end of discharge rates reported for basaltic volcanoes on land^{14–18}.

After the inflation/drain-back signal in the VSM2 pressure record, the instrument also recorded a further gradual subsidence of the sea floor by 1.4 m over the next 5 days (Fig. 3). This part of the signal is due to the deflation of the summit of Axial volcano as magma moved from the reservoir beneath the caldera and into the south rift zone. This deflation signal was also observed by two other sea-floor instruments during the same time period: VSM1, deployed near the centre of the caldera, recorded a subsidence of 3.2 m (ref. 5), and an array of extensometers on Axial's north rift zone measured a 4-cm decrease of a 405-m-long horizontal baseline⁶. This distance decrease can be explained if the baseline endpoints moved 68 cm radially inward toward the centre of the caldera during deflation⁶. These are, to our knowledge, the first *in situ* ground deformation measurements made during a submarine volcanic eruption. The VSM1 and extensometer data were used—before the data were recovered from VSM2—to calculate⁶ the depth of the magma reservoir (3.8 km) and the volume of magma removed from the reservoir beneath the caldera ($207 \times 10^6 \text{ m}^3$), on the basis of a point-source elastic deformation model¹⁹. We note that the third data point from VSM2 fits exactly on the same subsidence curve predicted by this deformation model (Fig. 3c). This agreement gives increased confidence in the model results, which are also consistent with estimates of the depth of Axial's magma reservoir based on seismic tomography²⁰.

Little is known about deep-sea eruptions, because we have only been able to detect them for about a decade and none have ever been witnessed. The data we report here, recorded by the VSM2 instrument caught in the 1998 lava flow at Axial volcano, were obtained by fortuitous circumstance. The instrument was simply in the right place at the right time, with the right sensors, and happened to have the right physical design to survive the eruption. □

Received 17 April; accepted 28 June 2001.

1. Fox, C. G., Dziak, R. P., Matsumoto, H. & Schreiner, A. E. Potential for monitoring low-level seismicity on the Juan de Fuca Ridge using military hydrophone arrays. *Mar. Technol. Soc. J.* **27**, 22–30 (1994).
2. Perfit, M. R. & Chadwick, W. W. Jr in *Faulting and Magmatism at Mid-Ocean Ridges* (eds Buck, W. R., Delaney, P. T., Karson, J. A. & Lagabriele, Y. 59–116 (AGU Geophysical Monograph 106, American Geophysical Union, Washington DC, 1998).
3. Dziak, R. P. & Fox, C. G. The January 1998 earthquake swarm at Axial Volcano, Juan de Fuca Ridge: Hydroacoustic evidence of seafloor volcanic activity. *Geophys. Res. Lett.* **26**, 3429–3432 (1999).
4. Embley, R. W., Chadwick, W. W. Jr, Clague, D. & Stakes, D. The 1998 eruption of Axial Volcano: Multibeam anomalies and seafloor observations. *Geophys. Res. Lett.* **26**, 3425–3428 (1999).
5. Fox, C. G. In situ ground deformation measurements from the summit of Axial Volcano during the 1998 volcanic episode. *Geophys. Res. Lett.* **26**, 3437–3440 (1999).
6. Chadwick, W. W. Jr, Embley, R. W., Milburn, H. B., Meinig, C. & Stapp, M. Evidence for deformation associated with the 1998 eruption of Axial Volcano, Juan de Fuca Ridge, from acoustic extensometer measurements. *Geophys. Res. Lett.* **26**, 3441–3444 (1999).
7. Baker, E. T., Fox, C. G. & Cowen, J. P. In situ observations of the onset of hydrothermal discharge during the 1998 submarine eruption of Axial Volcano, Juan de Fuca Ridge. *Geophys. Res. Lett.* **26**, 3445–3448 (1999).
8. Griffiths, R. W. & Fink, J. H. Solidification and morphology of submarine lavas: A dependence on extrusion rate. *J. Geophys. Res.* **97**, 19729–19737 (1992).
9. Gregg, T. K. P. & Fink, J. H. Quantification of submarine lava-flow morphology through analog experiments. *Geology* **23**, 73–76 (1995).
10. Gregg, T. K. P. & Chadwick, W. W. Jr. Submarine lava-flow inflation: A model for the formation of lava pillars. *Geology* **24**, 981–984 (1996).
11. Chadwick, W. W. Jr, Gregg, T. K. P. & Embley, R. W. Submarine lineated sheet flows: A unique lava morphology formed on subsiding lava ponds. *Bull. Volcanol.* **61**, 194–206 (1999).
12. Fox, C. G. Evidence of active ground deformation on the mid-ocean ridge: Axial Seamount, Juan de Fuca Ridge, April–June, 1988. *J. Geophys. Res.* **95**, 12813–12822 (1990).
13. Gregg, T. K. P. & Fornari, D. J. Long submarine lava flows: Observations and results from numerical modeling. *J. Geophys. Res.* **103**, 27517–27532 (1998).
14. Wadge, G. The variation of magma discharge during basaltic eruptions. *J. Volcanol. Geotherm. Res.* **11**, 139–168 (1981).
15. Lockwood, J. P. & Lipman, P. W. Holocene eruptive history of Mauna Loa Volcano. *Prof. Pap. US Geol. Surv.* **1350**, 509–536 (1987).
16. Wolfe, E. W., Neal, C. A., Banks, N. G. & Duggan, T. J. Geologic observations and chronology of eruptive events. *Prof. Pap. US Geol. Surv.* **1463**, 1–98 (1988).
17. Rowland, S. K. & Walker, G. P. L. Pahoehoe and a'a in Hawaii: Volumetric flow rate controls the lava structure. *Bull. Volcanol.* **52**, 615–628 (1990).

18. Thordarson, T. & Self, S. The Laki (Skaftár Fires) and Grímsvötn eruptions in 1783–1785. *Bull. Volcanol.* **55**, 233–263 (1993).
19. Mogi, K. Relations between the eruptions of various volcanoes and the deformation of the ground surfaces around them. *Bull. Earthquake Res. Inst. Univ. Tokyo* **36**, 99–134 (1958).
20. West, M. E., Menke, W. & Tolstoy, M. Massive magma reservoir beneath Axial volcano, Juan de Fuca ridge. *Nature* (submitted).

Acknowledgements

We thank H. Milburn, C. Meinig and P. McLain for helping to build, maintain and rescue the VSM2 instrument; the ROPOS group and the crews of the NOAA ship *Ron Brown* and RV *Thompson* for their help with successful recovery of VSM2; and S. Carbotte and J. Fink for comments on the manuscript. This work was supported by the NOAA Vents Program and the West Coast and Polar Regions Undersea Research Center.

Correspondence and requests for materials should be addressed to W.W.C. (e-mail: chadwick@pml.noaa.gov).

Spanwise flow and the attachment of the leading-edge vortex on insect wings

James M. Birch & Michael H. Dickinson

Department of Integrative Biology, University of California, Berkeley, California 94720, USA

The flow structure that is largely responsible for the good performance of insect wings has recently been identified as a leading-edge vortex^{1,2}. But because such vortices become detached from a wing in two-dimensional flow¹, an unknown mechanism must keep them attached to (three-dimensional) flapping wings. The current explanation, analogous to a mechanism operating on delta-wing aircraft, is that spanwise flow through a spiral vortex drains energy from the vortex core³. We have tested this hypothesis by systematically mapping the flow generated by a dynamically scaled model insect while simultaneously measuring the resulting aerodynamic forces. Here we report that, at the Reynolds numbers matching the flows relevant for most insects, flapping wings do not generate a spiral vortex akin to that produced by delta-wing aircraft. We also find that limiting spanwise flow with fences and edge baffles does not cause detachment of the leading-edge vortex. The data support an alternative hypothesis—that downward flow induced by tip vortices limits the growth of the leading-edge vortex.

The failure of conventional theory to explain the hovering flight of insects prompted the search for unsteady aerodynamic mechanisms that could account for the good performance of flapping wings. One such mechanism, dynamic stall, is manifest by the formation of a leading-edge vortex (LEV) on the top surface of the wing at high angles of attack^{1–5}. However, in both experimental and computational studies of two-dimensional (2D) wings started from rest, the LEV builds in strength until it detaches from the wing and is shed into the wake, being replaced by a trailing-edge vortex of the opposite sign^{6,7}. This pattern of vortex growth and shedding repeats, giving rise to a trail of counter-rotating vortices known as a Kármán street. In three-dimensional (3D) models of both hawkmoths (*Manduca sexta*) and fruitflies (*Drosophila melanogaster*), the LEV is not quickly shed, but rather remains attached to the wing throughout the stroke^{3,8,9}. There are currently two hypotheses to explain this prolonged attachment of the LEV in 3D flows. One possibility, analogous to a mechanism thought to stabilize an attached vortex on delta-wing aircraft such as Concorde¹⁰, is that base-to-tip spanwise flow in the form of a spiral vortex limits the

growth of leading-edge vorticity by removing energy from the vortex core^{3,11}. This model, based on flow visualizations of a dynamically scaled mechanical hawkmoth, has been supported by a computational fluid-mechanics simulation¹². An alternative hypothesis is that induced downward flow from tip vortices and wake vorticity reduces the effective angle of attack, attenuating the growth of the LEV. The wing then rotates, reverses direction, and sheds the accumulated vorticity before the LEV can enlarge to an unstable condition and detach^{13,14}.

We tested both these hypotheses using a dynamically scaled robotic model of a *Drosophila* wing flapping in mineral oil⁸ (see Methods). One wing of the model was equipped with a sensor to measure aerodynamic forces. In addition we characterized the flow around the wing using digital particle velocimetry (DPIV; see Fig. 1a and Methods). Flow visualization indicated that fluid separates at the leading edge of the wing, forming a shear layer of vorticity that concentrates into a large attached LEV (Fig. 1b). Velocity transects through the region of concentrated vorticity indicate a central core of rotational flow as expected of an idealized vortex¹⁵ (Fig. 1b, inset). Plots of spanwise velocity (u_z) reconstructed from rear views indicated that the axial flow velocity within the core of the LEV was quite small, only 2–5% of the average wing-tip velocity. Some axial flow within the core of the LEV is expected, driven by a spanwise pressure gradient that results from the variation in the

strength of vorticity along the length of the wing. In contrast to the low velocity of spanwise flow within the core, DPIV showed a broad sheet of base-to-tip spanwise flow on the dorsal surface of the rear two-thirds of the wing, with peak velocities approaching 40% of wing-tip velocity (Fig. 1c). The source of this flow is apparent when the wing was viewed from the rear (Fig. 1d). The laterally directed spanwise flow along the wing surface is induced by chord-wise vorticity, most of which rolls up into a coherent tip vortex that is contiguous with the LEV after it detaches from the wing and bends backward. In sum, the visualizations indicate that the flow structure generated by a model fruitfly wing at a Reynolds number of 160 is substantially different from that described for a model hawkmoth wing. We found no structure analogous to the spiral vortex of a delta-wing aircraft, and the core of the LEV, as defined by the region of rotation flow, contained axial flow of very low velocity.

Although the magnitude of axial flow within the core was small for the model fruitfly wing, it might still be critical for limiting the growth of the LEV and maintaining attachment throughout the stroke. To test this hypothesis, we constructed a wing with teardrop-shaped fences at $0.40R$ and $0.60R$ (where R is the distance from base to tip), which were designed to extend within the core region of the LEV and limit spanwise flow. When viewed from the rear, DPIV showed that the fences lowered the axial flow of fluid within the core below detectable levels, but had very little effect on the structure of

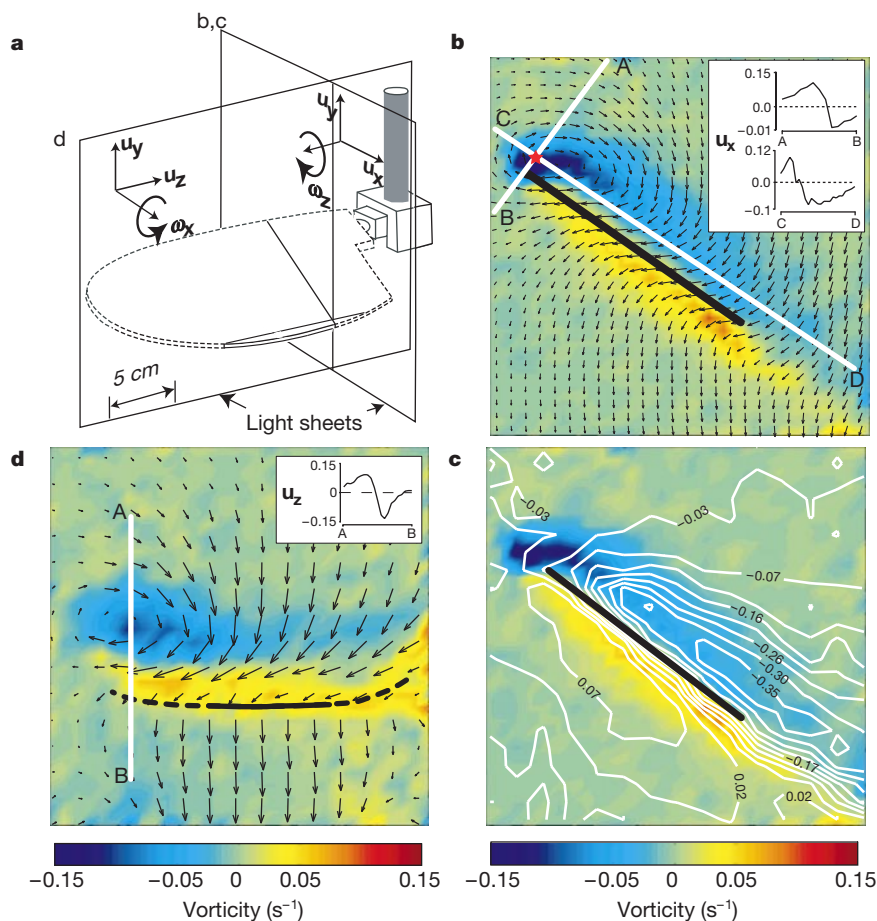


Figure 1 Maximum axial flow occurs behind the leading-edge vortex. **a**, Close-up of wing and the position of one light sheet during the separate 22 side-view and 34 rear-view digital particle image velocimetry (DPIV) experiments. **u**, velocity; **ω** , vorticity (subscripts **x**, **y** and **z** indicate vector components in Cartesian space). DPIV performed every 1 cm in both views. Letters indicate position of the following panels. **b**, Vorticity and flow pattern at $0.48R$ in chord-wise view at midstroke; leading edge to the left. Velocity vectors are superimposed on a 15×15 cm area around the wing. The black bar represents the wing,

red star represents location of maximum vorticity. White lines indicate u_x (m s^{-1}) transects graphed in inset. **c**, The spanwise velocity plot at $0.48R$ superimposed as a contour map on the vorticity plot. Values are normalized to instantaneous wing-tip velocity. Negative values represent flow out of the page. **d**, Rear view of vorticity at the trailing edge (black) with superimposed velocity vectors. Wing tip is to the left, and the white line represents the u_z (m s^{-1}) velocity transect (inset).

the LEV (Fig. 2a). This result, together with the absence of any structure resembling a spiral vortex, suggest that the maintenance of vortex attachment on fruitfly wings does not occur through a mechanism analogous to that operating on delta-wing aircraft.

However, one possibility is that the axial flow behind the leading edge plays the same role in removing energy from the growing vortex as the flow within a spiral vortex on a Concorde or hawkmoth. To test this hypothesis, we disrupted the broad posterior sheet of spanwise flow by constructing teardrop fences that projected rearward towards the trailing edge (Fig. 2b). But rather than increasing the strength of the LEV as expected if the spanwise flow limits vortex growth, this manipulation caused instead a small decrease in vortex strength as manifest by a 25% drop in net force. More importantly, as with the anteriorly directed fences, the rearward fences did not alter the time course of force generation relative to the fenceless wing, indicating that the experimental manipulations had no effect on the dynamics of vortex attachment throughout the stroke. In a third experiment aimed at manipulating

flow over the wing surface, we constructed a clear acrylic wall that exactly matched the sweep of the wing to block spanwise flow at the tip (Fig. 2c). As with the fences, this manipulation had no effect on the dynamics of vortex attachment. However, the presence of the cylindrical wall increased the total strength of the LEV by 14%, resulting in an 8% increase in force during the stroke. These results show that the LEV can remain attached throughout the stroke, with no signs of initiating Kármán shedding, even when its strength is artificially increased by a solid boundary at the tip. Thus, the net aerodynamic performance is limited by the magnitude of the LEV and not by its stable attachment throughout the stroke. Further, the flow structure generated by the fruitfly wing appears more robust than that described for a hawkmoth at higher Reynolds numbers⁴.

If spanwise flow does not remove energy from the LEV in 3D, what then explains its constant attachment relative to the 2D case? One possibility, suggested by the rise in vortex strength in the presence of a wall, is that downward flow induced by both the tip vortex and previously shed wake vorticity substantially lowers the effective angle of attack thereby limiting the growth of the LEV compared to the 2D case. On conventional fixed-wing aircraft, downwash from tip vortices reduces the 'effective' or 'aerodynamic' angle of attack, resulting in a phenomenon termed induced drag¹⁵. As it approaches the undersurface of the flapping wing, fluid from the free stream is gradually deflected downward until it is oriented parallel to the surface of the wing near the boundary layer (Fig. 3a). We estimated the aerodynamic angle of attack by measuring the mean orientation of flow vectors in the region between the free stream and the undersurface of the wing (box B, Fig. 3a). The aerodynamic angle of attack changes as a function of wing span from 43° at the tip to as low as 21° at 0.40R (Fig. 3b). Even in 2D, the angle of attack measured within this transition region would be lower than estimated from the free stream due to downwash induced by spanwise vorticity attached to the wing. However, the effect in 3D will be substantially worse due to the flow induced by chord-wise tip vorticity and the free vorticity within the wake shed during previous strokes.

In order to examine the influence of this additional downwash on force production, we examined the flows and forces generated by a sequence of successive downstrokes starting from rest (Fig. 3c–e). In the first stroke, when there is a tip vortex present but the wake is not yet established, both the lift and aerodynamic angle of attack are high (Fig. 3d, e). By the third stroke, the aerodynamic angle of attack and the mean lift reach an asymptotic level as the full wake structure becomes established. In the second stroke, the measured values for both the angle of attack and mean lift were markedly lower than in the first stroke or all subsequent strokes. One possible explanation for this transient attenuation, which we consistently observed, is that the deleterious influence of the wake is particularly strong in the second stroke because of the increased vorticity generated and shed during the first stroke. However, no matter what is responsible for especially large downwash in the second stroke, there is a strict correlation between the aerodynamic angles of attack and lift ($r^2 = 0.998$), which suggests that downwash has a potent inhibitory effect on the strength of the LEV and the magnitude of force production. In total, the flow visualizations and force measurements suggest that the downward flow induced by the tip vortex and wake vorticity limit the growth of the LEV and contribute to its prolonged attachment.

Because a spiral vortex is clearly present on a robotic model hawkmoth, our results suggest that the precise flow structure of a LEV may depend critically on Reynolds number. Fruitflies flap their wings within a range of Reynolds number of 100–250, whereas the much larger hawkmoth operates at Reynolds numbers in excess of 2,000. One possible explanation for the difference in flow structure is that the pressure gradients within the vortex core are too small to drive a substantial axial flow on the smaller wing of *Drosophila*.

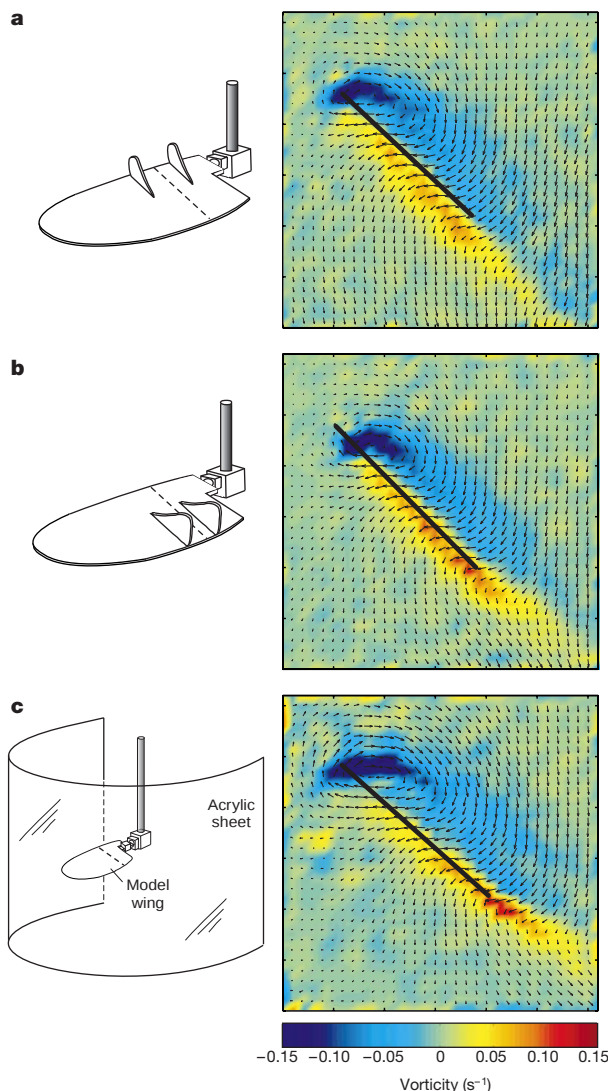


Figure 2 The leading edge vortex (LEV) remains attached despite experimental manipulation. Left column shows wing morphology, dotted line is position ($0.48R$) of vorticity plots (right column). **a**, Forward-pointing fences. **b**, Rearward-pointing fences. **c**, Wall. Compared to the two fence manipulations, the presence of a wall causes the LEV to enlarge and extend farther along the wing before detaching into a tip vortex.

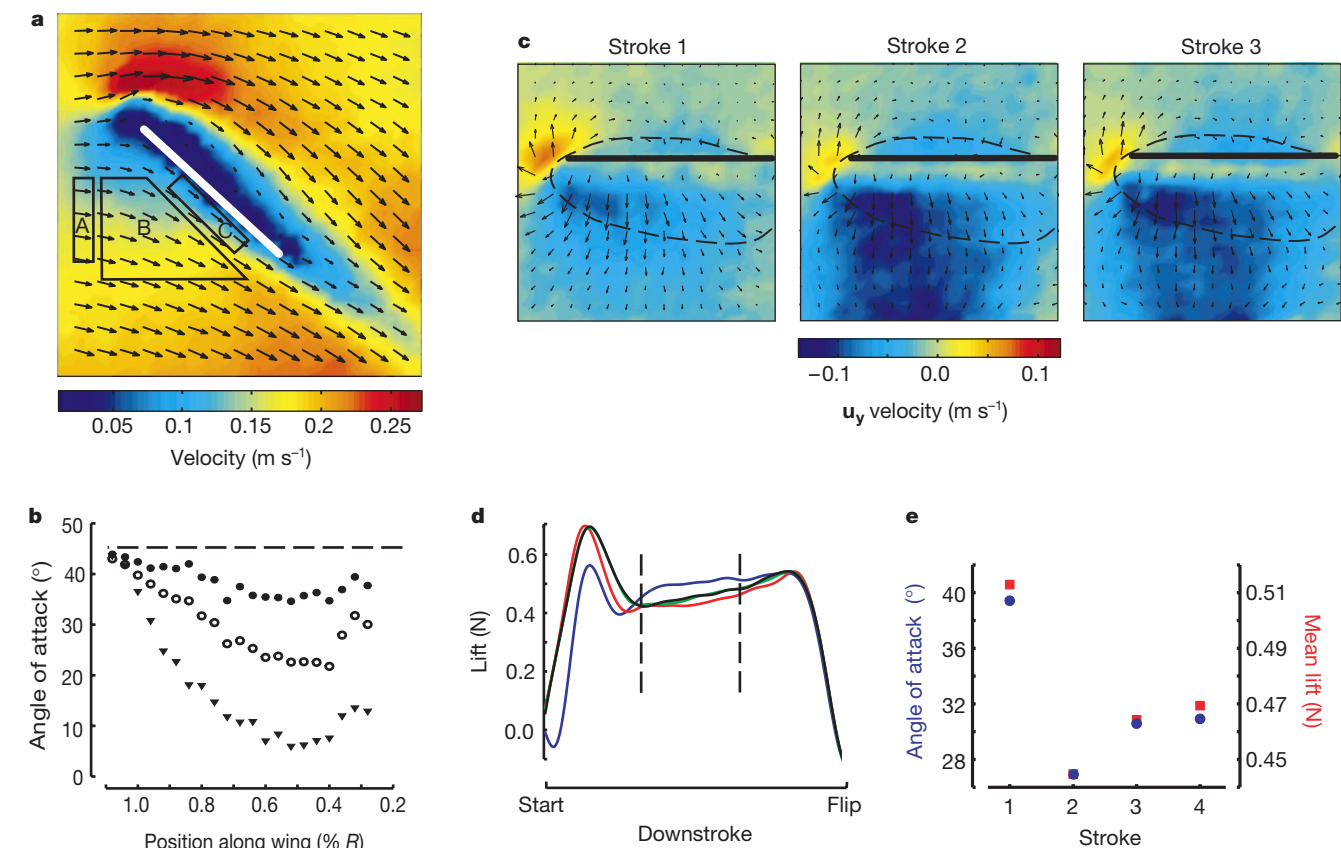


Figure 3 Induced downwash lowers both the aerodynamic angle of attack and the lift generated by the LEV. **a**, Wing viewed from the side (white bar) at $0.60R$. Shown are flow velocity magnitudes (pseudocolour) with velocity vectors superimposed during a fourth stroke. We calculated effective angle of attack in three areas along the wing (boxes A, B and C). **b**, Although the wing was set to an angle of attack of 45° throughout the stroke, the effective angle of attack changes along the wing from tip to base; closed circles represent mean values from box A; open circles, box B; triangles, box C. **c**, Pseudocolour plots represent vertical fluid velocity when viewing wing from the rear, 2 cm behind the

leading edge. Mean U_y velocity within box B from (a): stroke 1, 0.0026 m s^{-1} ; stroke 2, 0.0056 m s^{-1} ; stroke 3, 0.0054 m s^{-1} . **d**, Lift forces during four downstrokes. Note the attenuating effect of the wake on forces during translation (middle third of stroke). First stroke, blue; second stroke, red; third stroke, green; fourth stroke, black. **e**, Mean effective angle of attack measured in box B (blue) and mean lift force (red) for the central third of the wing and central third of the stroke (dashed lines in **d**), respectively, for each of four strokes.

However, because the body length of a typical adult insect is in the range of 4–5 mm (refs 16, 17), fruitflies (body length ~ 2.5 –4 mm) might represent a better general model for the fluid mechanics of hovering flight than the much larger hawkmoths (body length ~ 42 mm; ref. 18). In addition, the observation that a flapping wing at lower Reynolds number can generate comparable performance in the absence of a spiral vortex raises the possibility that while such a structure is present at higher Reynolds numbers, it is not necessary for the continued attachment of the LEV. We suggest instead that the prolonged attachment of the LEV on insect wings may be due to the attenuating effect of the downwash induced by wake vorticity. □

Methods

Robotic model

This consists of two robotic arms immersed in a $1 \times 1 \times 2 \text{ m}$ Plexiglas tank filled with mineral oil (density, $0.88 \times 10^3 \text{ kg m}^{-3}$; kinematic viscosity, 115 cSt)⁸. In this experiment, we removed one arm and describe the pattern from a single flapping wing. We constructed a model *Drosophila* wing from 2.25-mm-thick clear acrylic, and flapped it at 168 mHz in a symmetric pattern (stroke amplitude, 180° ; angle of attack at midstroke, 45° up and downstroke), which resulted in a mean Reynolds number of 160. A force sensor at the wing base measured forces orthogonal and parallel to the wing surface, from which we calculated lift and drag coefficients. To limit axial flow, we built two wings with teardrop fences at $0.40R$ and $0.60R$. One wing had fences extending toward the leading edge, the other wing had fences toward the trailing edge. We also built an acrylic wall with radius equal to wing length (25 cm). When in position, this wall (extending 21 cm above and 35 cm below the wing) covered 210° of arc, and was within 1 mm of the wingtip throughout the stroke.

Digital particle image velocimetry

To visualize wake structure, the tank of mineral oil was filled with bubbles by pumping air through a ceramic water-purifier filter. After larger bubbles had risen to the top of the tank, smaller bubbles were effectively stationary relative to wing motion and these created a highly concentrated and stable seed for DPIV. We used a dual Nd:YAG laser system (Insight v. 3.0, TSI) to create two identically positioned light sheets $\sim 2.5 \text{ mm}$ thick separated by $2,500 \mu\text{s}$. A two-frame cross-correlation of pixel intensity peaks with 50% overlap of 64×64 pixel interrogation areas resulted in 900 velocity vectors per sheet position. Whether near the clear wing or in the fluid, vector validation resulted in removal of an average of 1.2 vectors per slice (less than 0.2% of vectors in each image); these missing values were filled by interpolation from a 3×3 nearest neighbour matrix. For each wing (regular, front-fence, rear-fence, wall) we acquired 24 views, each $20 \times 20 \text{ cm}$, perpendicular to the long axis of the wing (that is, from the side) at mid-downstroke, moving out the wing from $0.24R$ to $1.12R$ (where R is wing length) in 1-cm increments. We then moved the laser and camera to acquire 34 views parallel to the long axis of the wing (that is, from the rear), starting 6 cm in front of the leading edge and moving in 1-cm increments to 14 cm behind the trailing edge of the wing. From these 34 rear views, we constructed new arrays from the wing base to tip that held y - z velocity and x vorticity. Although the morphological angle of attack (AOA) was 45° , we calculated the effective AOA (α') at each point in the fluid as $\alpha' = 45 - \tan^{-1}(u_y/u_x)$. All calculations were made using custom software programs written in Matlab (v.5.3, MathWorks).

Received 11 October 2000; accepted 17 May 2001.

- Dickinson, M. H. & Götz, K. G. Unsteady aerodynamic performance on model wings at low Reynolds numbers. *J. Exp. Biol.* **174**, 45–64 (1993).
- Ellington, C. P. The aerodynamics of hovering insect flight. IV. Aerodynamic mechanisms. *Phil. Trans. R. Soc. Lond. B* **305**, 79–113 (1984).
- Ellington, C. P., Van den Berg, C., Willmott, A. P. & Thomas, A. L. R. Leading-edge vortices in insect flight. *Nature* **384**, 626–630 (1996).
- Van den Berg, C. & Ellington, C. P. The three-dimensional leading-edge vortex of a 'hovering' model hawkmoth. *Phil. Trans. R. Soc. Lond. B* **352**, 329–340 (1997).

5. Sunada, S., Sakaguchi, A. & Kawachi, K. Airfoil section characteristics at low Reynolds number. *J. Fluids Eng.* **119**, 129–135 (1997).
6. Kyja, M. & Arie, M. A contribution to an inviscid vortex-shedding model for an inclined flat plate in uniform flow. *J. Fluid Mech.* **82**, 223–240 (1977).
7. Sarpkaya, T. An inviscid model of two-dimensional vortex shedding for transient and asymptotically steady separated flow over an inclined plane. *J. Fluid Mech.* **68**, 109–128 (1975).
8. Dickinson, M. H., Lehmann, F. O. & Sane, S. P. Wing rotation and the aerodynamic basis of insect flight. *Science* **284**, 1954–1960 (1999).
9. Willmott, A. P., Ellington, C. P. & Thomas, A. L. R. Flow visualization and unsteady aerodynamics in the flight of the hawkmoth *Manduca sexta*. *Phil. Trans. R. Soc. Lond. B* **352**, 303–316 (1997).
10. Kücheman, D. *The Aerodynamic Design of Aircraft* (Pergamon, Oxford, 1978).
11. Maxworthy, T. Experiments on the Weis-Fogh mechanism of lift generation by insects in hovering flight Part 1. Dynamics of the 'fling'. *J. Fluid Mech.* **93**, 47–63 (1979).
12. Liu, H., Ellington, C. P., Kawachi, K., Van den Berg, C. & Willmott, A. A computational fluid dynamic study of hawkmoth hovering. *J. Exp. Biol.* **201**, 461–477 (1998).
13. Dickinson, M. H. The effects of wing rotation on unsteady aerodynamic performance at low Reynolds numbers. *J. Exp. Biol.* **192**, 179–206 (1994).
14. Wang, J. Two dimensional mechanism for insect hovering. *Phys. Rev. Lett.* **85**, 2216–2219 (2000).
15. Milne-Thomson, L. M. *Theoretical Aerodynamics* (Macmillan, New York, 1966).
16. May, R. M. in *Diversity of Insect Faunas* (eds Mound, L. A. & Waloff, N.) 188–204 (Blackwell Scientific, London, 1978).
17. Dudley, R. *The Biomechanics of Insect Flight. Form, Function, Evolution* (Princeton Univ. Press, Princeton, 2000).
18. Willmott, A. P. & Ellington, C. P. The mechanics of flight in the hawkmoth *Manduca sexta*. II. Aerodynamic consequences of kinematic and morphological variation. *J. Exp. Biol.* **200**, 2723–2745 (1997).

Acknowledgements

We thank C. Ellington, M. Gahrib, G. Lauder, S. Sane and J. Wang for comments and suggestions on this manuscript. This work was supported by the National Science Foundation, Office of Naval Research, and DARPA.

Correspondence and requests for materials should be addressed to M.H.D. (e-mail: flymanmd@socrates.berkeley.edu).

Change in pattern of ongoing cortical activity with auditory category learning

F. W. Ohl^{1,†}, H. Scheich^{1,*} & W. J. Freeman^{2,†}

¹Leibniz-Institut für Neurobiologie, Brenneckstraße 6, D-39118 Magdeburg, Germany

²Department of Molecular and Cell Biology, University of California, 129 LSA, Berkeley 94720-3200, USA

Humans are able to classify novel items correctly by category^{1,2}; some other animals have also been shown to do this^{3–7}. During category learning, humans group perceptual stimuli by abstracting qualities from similarity relationships of their physical properties^{1,2,8}. Forming categories is fundamental to cognition⁹ and can be independent of a 'memory store' of information about the items or a prototype¹⁰. The neurophysiological mechanisms underlying the formation of categories are unknown. Using an animal model of category learning⁶, in which frequency-modulated tones are distinguished into the categories of 'rising' and 'falling' modulation, we demonstrate here that the sorting of stimuli into these categories emerges as a sudden change in an animal's learning strategy. Electro-corticographic recording from the auditory cortex¹¹ shows that the transition is accompanied by a change in the dynamics of cortical stimulus representation. We suggest that this dynamic change represents a mechanism underlying the recognition of the abstract quality (or qualities) that defines the categories.

Mongolian gerbils (*Meriones unguiculatus*) can be trained with pairs of linearly rising and falling frequency-modulated tones (Fig. 1a) to establish the categories of 'rising' and 'falling' modula-

tion and transfer the concept of modulation direction to novel stimuli⁶. Transfer to novel stimuli is considered to be the most decisive behavioural index for category learning^{12,13}. The training consists of six sequential 'training blocks' (see Methods), in each of which a different pair of rising and falling frequency-modulated tones is presented (Fig. 1a, numbers). In the early blocks, discrimination performance gradually improves in each block as the animals learn to discriminate between the rising and falling tones over a series of daily training sessions (discrimination learning; Fig. 1b). A sudden transition in behaviour then occurs, after which novel stimuli are immediately identified as belonging to their correct category: discrimination performance is already high in the first session of a training block, even though the stimulus pair to be discriminated is novel (categorization; Figs 1c, 2). The point during the training when this transition occurs differs between animals (Fig. 2), although in the 10 animals studied it was always after the second training block.

In addition to the sudden emergence of concept transfer we find a sudden alteration in the perceptual scaling characteristics of the animals: measuring psychometric functions for frequency-modulation rate after each training block shows a generalization gradient for modulation rate during the discrimination-learning phase; that is, the greater the difference in modulation rates (rates of change of frequency) between the training stimuli used in the training block and the test stimuli, the smaller the conditioned response rate

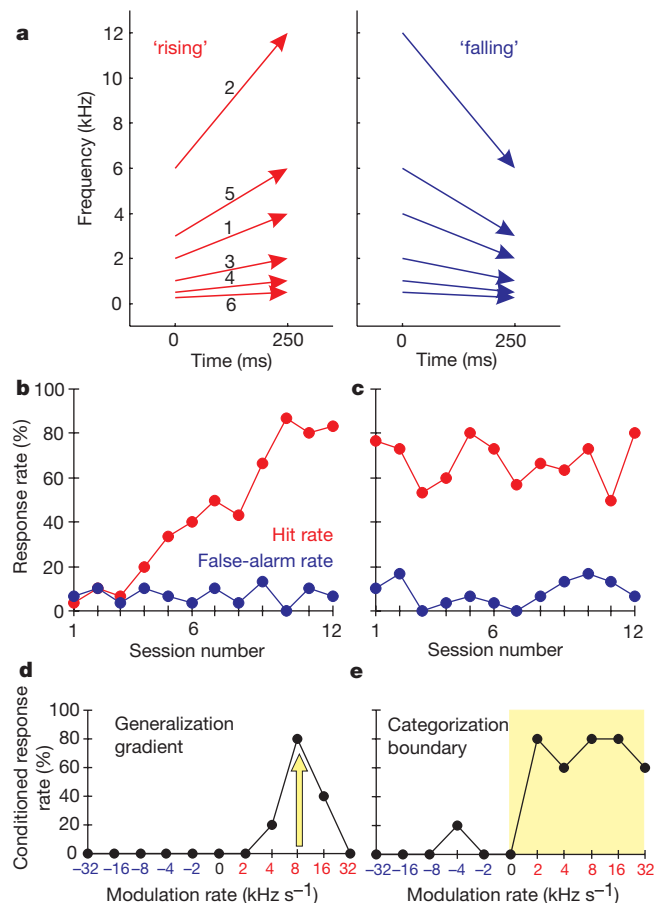


Figure 1 Stimuli and behavioural measures of category learning. **a**, Rising (red) and falling (blue) frequency-modulated tones used in the six sequential training blocks (numbers). **b**, Sample learning curve of gerbil 3 before transition to categorization. **c**, Sample learning curve of same animal after transition to categorization. **d**, Psychometric function for modulation rate obtained after training block shown in **b**. Peak modulation rate of 8 kHz s⁻¹ (arrow) corresponds to modulation from 2 kHz to 4 kHz in 250 ms used in this block. **e**, Sigmoid psychometric function obtained after training block shown in **c**.

(Fig. 1d). After the transition, the psychometric function becomes sigmoid (categorical perception, Fig. 1e), indicating that the modulation rate is being perceived as a member either of the 'rising' or the 'falling' category.

The development of spatiotemporal activity patterns in primary auditory cortex was studied in four animals using high-resolution surface electrocorticography¹¹ during the training, including the transition from discrimination learning to categorization. Record-

ings from four initially naive gerbils were made with an 18-electrode array on the surface of the right primary auditory cortex (Fig. 3a), which contains neurons performing the high degree of spectral integration^{14,15} that is essential for the discrimination of frequency-modulated tones¹⁶. Sample recordings are shown in Fig. 3b. This recording technique provides sufficient resolution to detect differences in spatial activity patterns elicited by pure tones in a frequency range comparable to that traversed by the frequency-modulated tones¹¹.

Electrophysiological correlates of category learning do not necessarily occur at times locked to the presentation of a stimulus, so

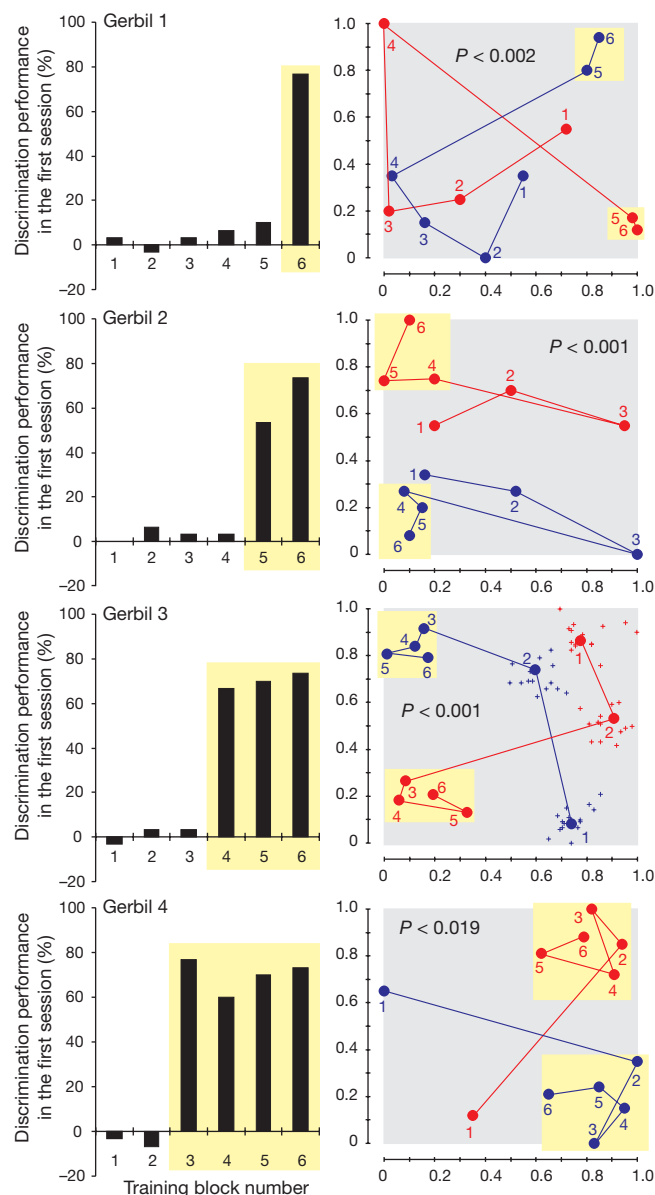


Figure 2 Behavioural transition to categorization (left column) parallels development of cortical spatial activity patterns (right column). Left column: discrimination performance in the first session of each of the six training blocks. Yellow areas indicate categorization phase (Wilcoxon's test, $P < 0.05$). Right column: similarity relations between spatial activity patterns during the marked states (compare with Fig. 3d). Transition to category learning in the behavioural data correlates with clustering (P -values of resampling test given) of the marked states 'within category' (yellow areas). Only the activity pattern during the marked state that gave rise to the maximum peak value of the dissimilarity function for each category is plotted for each training block (numbers). For gerbil 3, marked states of later sessions in blocks 1 and 2 have been included (+) to demonstrate that these point clouds do not fall into the clusters found after the transition to categorization. Absolute coordinates of points have no particular meaning other than scaling relative distances between any pair of points.

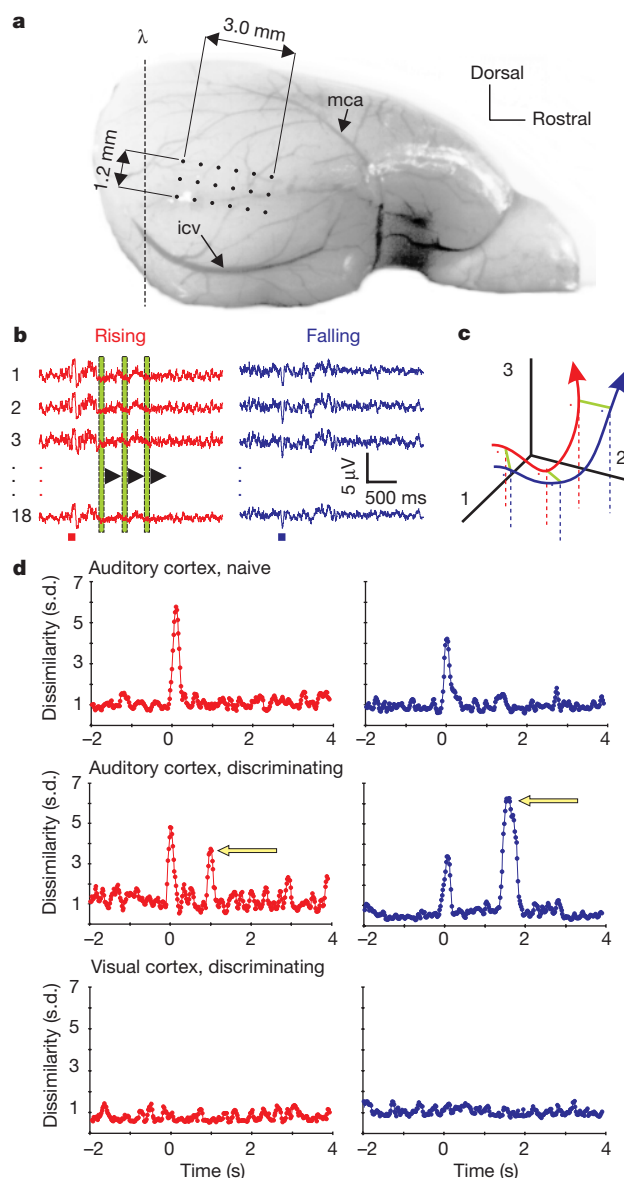


Figure 3 Measurement and analysis of ongoing spatiotemporal activity. **a**, Positioning of electrode array over right auditory cortex. λ , caudal fusion point of parietal bones; icv, inferior cerebral vein; mca, middle cerebral artery. **b**, Sample surface electrocorticograms from a single trial using a rising (red) and a falling (blue) frequency-modulated tone. Stimulus markers (squares) are indicated below bottom traces. Green bars indicate the sample time windows used to calculate the r.m.s. amplitudes. **c**, Mapping spatial patterns of 18 r.m.s. amplitudes to points in 18-dimensional state space (three dimensions shown). Dissimilarity between spatial patterns can be quantified by euclidean distances between corresponding trajectory points (green lines). **d**, Sample 'dissimilarity functions' calculated between a single trajectory and the average over all trajectories of the opposite stimulus category. Marked states are indicated with arrows.

classical averaging techniques cannot be applied in the analysis of the electrocorticograms. Instead, we developed an analysis that can identify neural activity patterns of putative relevance to the coding of category-specific information in single trials independent of their locking to the stimulus. The spatiotemporal activity pattern in each 'rising' trial was compared to a spatiotemporal reference pattern given by the mean across all 'falling' trials within a training session; the reverse comparison was made for all 'falling' trials. The dissimilarity in spatial pattern between the investigated trial and the reference pattern, termed the 'dissimilarity function', was determined as a function of time for each comparison (Fig. 3c).

Analysis of the 2 s before stimulus presentation revealed a baseline to which post-stimulus responses can be related. After stimulus onset, the dissimilarity function peaked from baseline, indicating a transient activity state in which the dissimilarity function was locally maximum compared with the reference pattern. This peak, which reflects the stimulus perturbing the ongoing activity^{17,18}, is found in all trials even in naive animals (Fig. 3d, top row), but not in animals aroused by presentation of the unconditioned stimulus only. Additional peaks (1–5) emerged in each dissimilarity function (mean of 3.5 ± 1.3 peaks per trial) with continued training but not in the sensitization control (0.3 ± 0.5 peaks per trial) (Fig. 3d). The peaks were randomly distributed during the 4-s observation interval between stimulus onset and conditioned response, without apparent correlation to the response latency (coefficient of determination, $R^2 = 0.13$). These additional peaks tag transient states in the ongoing cortical activity, here referred to as 'marked states'. To further rule out contribution to the emergence of the marked states by factors unspecific to the auditory stimulus (for example, movement, response preparation), we implanted three gerbils with the same recording array over the visual cortex and trained them in the auditory model. We found neither the early peak nor any later marked states (0.3 ± 0.6 per trial) emerging in the activity of the visual cortex (Fig. 3d).

We tested whether category-specific information is processed during these marked states by analysing the similarities among the spatial activity patterns during the marked states in single trials. This showed that the behavioural transition to categorization (Fig. 2, left column) was reflected physiologically by changes in the cortical spatial activity patterns (Fig. 2, right column). Specifically, with categorization, the patterns reflected the 'belongingness' to the formed category rather than the physical characteristics of the stimuli. The similarity relations between spatial activity patterns during the marked states is depicted in two-dimensional plots (Fig. 2, right column), in which the dissimilarity between two patterns is proportional to the distance between the corresponding points. During discrimination learning, dissimilarities within and between categories were similar in magnitude. In contrast, the dissimilarities after the transition to categorization were significantly smaller within a category than between categories in all animals. For example, for Gerbil 2, which made the transition after training block 4, the points representing marked states in blocks 5 and 6 (after the transition) are much closer to that of block 4 in same category (yellow areas in plot) than to points 5 and 6 in the other category. In comparison, in blocks 1–4 (before the transition), points within a category are about as far apart as those between categories. For all four animals, we found such clustering of marked states within categories emerging with the behavioural transition to categorization (Fig. 2, right column).

Thus, two types of transient activity state discriminating between rising and falling frequency-modulated tones coexist in the same neuronal substrate. First, independent of the training state, an activity state occurred during a 150–250-ms time interval locked to stimulus onset, during which rising and falling frequency-modulated tones give rise to spatially distinct patterns in the tonotopic map, as was expected from previous frequency-modulated tone studies^{19,20}. The stimulus specificity of these activity states

therefore reflects the tonotopic organization of the early evoked cortical activity¹¹. Second, the transient states that emerge later in the 4-s interval between stimulus onset and reinforcement, which did not occur in the naive animal, are likely to be related to learning, particularly as their pattern reflected the learned category structure.

Similar learning-related spatial activity patterns have been reported from various cortical areas but their interpretation as sensory representations has posed a problem, because they typically do not show an invariant relationship to the trained stimuli^{21–24}. Our discovery that these states possess an internal metric, specific to the individual, that reflects the category structure rather than the physical features of the stimulus may resolve this issue. This relationship cannot be detected in a pure discrimination experiment or a contingency-reversal control. The category-learning experiment, however, provides an objective way to observe a subjectively generated cognitive structure. Furthermore, category-specific deficits observed clinically after circumscribed cortical lesions^{25–28} fits with our demonstration that category formation relies on a spatial representation in the cortex. □

Methods

Subjects and electrophysiological recording

Four male Mongolian gerbils (85–115 g) were chronically implanted with a 3×6 array of stainless steel microwires (100 μm diameter, 600 μm interelectrode distance) placed on the epidural surface over the right auditory cortex and centred over the primary field AI as judged by a preceding analysis of evoked potential topography¹¹. For control purposes three animals were implanted with the array over the right visual cortex. During the behavioural experiments, monopolar recordings against a fronto-parietal reference were made from the each of the 18 electrodes, amplified (10,000 $\times$), filtered (fall-off: 6 dB octave⁻¹, 3-dB cutoff frequencies: 0.1 Hz, 100 Hz), and stored on a computer for off-line analysis.

Stimuli and behavioural procedures

Stimuli consisted of linearly frequency-modulated tones (250 ms duration, 5 ms linear onset and offset ramps, modulation ranges see Fig. 1a) delivered in free field through calibrated loudspeakers at 70 dB sound pressure level. Using a previously designed model, we trained animals to form the concept of modulation direction and establish the categories of 'rising' and 'falling' frequency-modulated tones⁶. Experiments took place in a two-compartment shuttle box (E10-15, Coulbourn) positioned in a sound-attenuating Faraday cage. In each trial, the animals received either a rising or a falling tone and were trained to discriminate between them in a GO/(NO-GO) procedure motivated by negative reinforcement (150–300 μA , electrodermal stimulation administered through a metal floor grid) after misses (failing to GO) or false alarms (mistaken GO). The GO-conditioned response, that is, moving to the other compartment of the shuttle box, had to be accomplished within 4 s after tone onset to count as a hit trial. Training was given in six sequential blocks with daily sessions of 30 trials using rising frequency-modulated tones and 30 trials using falling frequency-modulated tones in a randomized sequence. In each block, one of the frequency-modulated tone pairs (Fig. 1a) was used and training continued until significant discrimination was observed for at least four consecutive sessions. When each training block was completed, psychometric functions for modulation rate were measured following a standard procedure²⁹ (see Supplementary Information). Modulation rates varied in steps of ± 2 octaves s^{-1} with respect to the modulation rate of the stimulus pair used in the training block. These measurements yielded generalization gradients in the discrimination phase and categorical psychometric functions in the categorization phase.

Discrimination performance in each session was quantified by the difference between hit rate and false-alarm rate, expressed as a percentage, and tested by Wilcoxon's test against the null hypothesis of equal rates, using a criterion of $P < 0.05$ for significant discrimination. Transition from discrimination learning to categorization was indicated by the instantaneous transfer of the concept of modulation direction to a novel frequency-modulated tone pair, which showed as a high-discrimination performance for the novel stimuli early in the first training session of a block (see Fig. 2). To test the specificity of electrophysiological correlates of learning, we carried out an arousal control (stimulation with the unconditioned stimulus alone and a sensitization control (60 trials of uncorrelated presentations of conditioned stimulus and unconditioned stimulus with each animal before onset of the described categorization training).

Analysis of electrophysiological data

In each trial the data from the 18-channel electrocorticogram was analysed in a 6-s time interval (2 s before stimulus + 4 s after stimulus). For each recording channel a 120-ms time window was moved through the data in steps of 20 ms and the mean signal power in each window was determined by calculating the root-mean-square (r.m.s.) amplitude. The power values obtained from the 18 recording channels in each window were summarized into an 18-dimensional state vector that characterized the spatial distribution of activity in that window (Fig. 3b). Each vector was z-transformed to remove activity common to the channels. The temporal development of the spatial activity pattern corresponded to a trajectory described by the state vector in an abstract 18-dimensional configuration space

as the window was stepped along the time axis (Fig. 3c). We calculated the distance in configuration space between the trajectory for each trial (the presentation of either a rising or a falling frequency-modulated tone) and a trajectory formed by averaging over all trajectories for the other stimulus in that session; that is, a single trial with a rising frequency-modulated tone was compared with the mean over all trials with falling frequency-modulated tones and vice versa. The distance is a dissimilarity function, a measure of the dissimilarity between the compared spatial activity patterns expressed in standard deviations (s.d.) of the mean dissimilarity obtained over the entire trajectory (Fig. 3d). Marked states are defined as peaks in the dissimilarity function with amplitudes at least three s.d. over the baseline of the amplitude distribution. We used a nonlinear projection algorithm³⁰ to display the combinatorial multitude of similarity relations between marked states, resulting in a map of the corresponding 18-dimensional state vectors in the two-dimensional plane that preserves all mutual distances between projected states. This is achieved by minimizing the relative mapping error

$$E(k) = \frac{1}{\Delta} \sum_{i,j=1}^N \frac{(d_{ij}^* - d_{ij}(k))^2}{d_{ij}^*}$$

where

$$\Delta = \sum_{i,j=1}^N d_{ij}^*$$

and d_{ij}^* is the euclidean distance between marked state vectors i and j in the original 18-dimensional state space, $d_{ij}(k)$ is the distance between the corresponding vectors projected into the two-dimensional plane after the k th step, and N is the number of vectors projected. The mapping error was minimized with a step-wise steepest-gradient procedure starting with a two-dimensional uniformly random vector distribution in the unit square. The mapping was accepted when the criterion

$$E(k \rightarrow \infty) \leq 0.5\%$$

was reached; when this was not achieved, another random starting condition for the steepest-gradient procedure was used. Clustering of the spatial patterns during the marked states after categorization (Fig. 2, right column) was tested using a resampling approach (see Supplementary Information).

Received 12 April; accepted 24 May 2001.

- Komatsu, L. K. Recent views of conceptual structure. *Psychol. Bull.* **112**, 500–526 (1992).
- Medin, D. L. & Coley, J. D. in *Perception and Cognition at Century's End: History, Philosophy, Theory (Handbook of Perception and Cognition)* (eds Hochberg, J. & Cutting, J. E.) 403–432 (Academic, New York, 1998).
- Woodruff, G. & Premack, D. Primitive mathematical concepts in the chimpanzee: proportionality and numerosity. *Nature* **293**, 568–570 (1981).
- Herrnstein, R. J. Levels of stimulus control. A functional approach. *Cognition* **37**, 133–166 (1990).
- Pearce, J. M. in *Current Topics in Animal Learning: Brain, Emotion, and Cognition* (eds Dachowski, L. & Flaherty, C.) 141–164 (Erlbaum, Hillsdale, 1991).
- Wetzel, W., Wagner, T., Ohl, F. W. & Scheich, H. Categorical discrimination of direction in frequency-modulated tones by Mongolian gerbils. *Behav. Brain Res.* **91**, 29–39 (1998).
- Orlov, T., Yakovlev, V., Hochstein, S. & Zohary, E. Macaque monkeys categorize images by their ordinal number. *Nature* **404**, 77–80 (2000).
- Feldman, J. Minimization of Boolean complexity in human concept learning. *Nature* **407**, 630–633 (2000).
- Smith, E. E. & Jonides, J. in *The New Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 1013–1022 (MIT Press, Cambridge, Massachusetts, 2000).
- Knowlton, B. J. & Squire, L. R. The learning of categories: parallel brain systems for item memory and category knowledge. *Science* **262**, 1747–1749 (1993).
- Ohl, F. W., Scheich, H. & Freeman, W. J. Topographic analysis of epidural pure-tone-evoked potentials in gerbil auditory cortex. *J. Neurophysiol.* **83**, 3123–3132 (2000).
- Lea, S. E. G. in *Animal Cognition* (eds Roitblat, H. L., Bever, T. G. & Terrace, H. S.) 263–276 (Erlbaum, Hillsdale, 1984).
- Moody, D. B. in *Methods in Comparative Psychoacoustics* (eds Klump, G. M., Dooling, R. J., Fay, R. R. & Stebbins, W. C.) 293–305 (Birkhäuser, Basel, 1995).
- Schreiner, C. E., Read, H. L. & Sutter, M. L. Modular organization of frequency integration in primary auditory cortex. *Annu. Rev. Neurosci.* **23**, 501–529 (2000).
- Ohl, F. W. & Scheich, H. Orderly cortical representation of vowels based on formant interaction. *Proc. Natl Acad. Sci. USA* **94**, 9440–9444 (1997).
- Ohl, F. W., Wetzel, W., Wagner, T., Rech, A. & Scheich, H. Bilateral ablation of auditory cortex in Mongolian gerbil affects discrimination of frequency modulated tones but not of pure tones. *Learn. Mem.* **6**, 347–362 (1999).
- Freeman, W. J. Neural mechanisms underlying destabilization of cortex by sensory input. *Physica D* **75**, 151–164 (1994).
- Arieli, A., Sterkin, A., Grinvald, A. & Aertsen, A. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).
- Heil, P. & Scheich, H. Spatial representation of frequency-modulated signals in the tonotopically organized auditory cortex analogue of the chick. *J. Comp. Neurol.* **322**, 548–565 (1992).
- Huottilainen, M. et al. Sustained fields of tones and glides reflect tonotopy of the auditory cortex. *NeuroReport* **6**, 841–844 (1995).
- Freeman, W. J. & Schneider, W. Changes in spatial patterns of rabbit olfactory EEG with conditioning to odors. *Psychophysiology* **19**, 44–56 (1982).
- Freeman, W. J. & Grajski, K. A. Relation of olfactory EEG to behavior: factor analysis. *Behav. Neurosci.* **101**, 766–777 (1987).

- Barrie, J. M., Freeman, W. J. & Lenhart, M. Modulation by discriminative training of spatial patterns of gamma EEG amplitude and phase in neocortex of rabbits. *J. Neurophysiol.* **76**, 520–539 (1996).
- Freeman, W. J. *Neurodynamics: an Exploration in Mesoscopic Brain Dynamics*. (Springer, London, 2000).
- Hart, J. Jr, Berndt, R. S. & Caramazza, A. Category-specific naming deficit following cerebral infarction. *Nature* **316**, 439–440 (1985).
- McCarthy, R. A. & Warrington, E. K. Evidence for modality-specific meaning systems in the brain. *Nature* **334**, 428–430 (1988).
- Creutzfeldt, O., Ojemann, G. & Lettich, E. Neuronal activity in human lateral temporal lobe. I. Responses to speech. *Exp. Brain Res.* **77**, 451–475 (1989).
- Martin, A., Wiggs, C. L., Ungerleider, L. G. & Haxby, J. V. Neural correlates of category-specific knowledge. *Nature* **379**, 649–652 (1996).
- Kuhl, P. K. & Miller, J. D. Speech perception by the chinchilla: voiced–voiceless distinction in alveolar plosive consonants. *Science* **190**, 69–72 (1975).
- Sammon, J. W. Jr A nonlinear mapping for data structure analysis. *IEEE Transact. Comp.* **C-18**, 401–409 (1969).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank K. Buckisch, B. Burke, M. Deliano and D. Labra-Cardero for technical assistance. We also thank J. Altman for critical comments on an earlier version of the manuscript.

Correspondence and requests for materials should be addressed to F.W.O. (e-mail: frank.ohl@ifn-magdeburg.de).

Purification of a pluripotent neural stem cell from the adult mouse brain

Rodney L. Rietze*, Helen Valcanis†, Gordon F. Brooker*, Tim Thomas*, Anne K. Voss* & Perry F. Bartlett*

* The Walter and Eliza Hall Institute of Medical Research, Royal Parade, Parkville, Victoria 3050, Australia

† Howard Florey Institute, University of Melbourne, Parkville, Victoria 3010, Australia

The adult mammalian central nervous system (CNS) contains a population of neural stem cells (NSCs)^{1–4} with properties said to include the generation of non-neural progeny^{5–7}. However, the precise identity, location and potential of the NSC *in situ* remain unclear. We purified NSCs from the adult mouse brain by flow cytometry, and directly examined the cells' properties. Here we show that one type of NSC, which expresses the protein nestin but only low levels of PNA-binding and HSA proteins, is found in both ependymal and subventricular zones and accounts for about 63% of the total NSC activity. Furthermore, the selective depletion of the population of this stem cell in *querkopf*⁸ mutant mice (which are deficient in production of olfactory neurons) suggests that it acts as a major functional stem cell *in vivo*. Most freshly isolated NSCs, when co-cultured with a muscle cell line, rapidly differentiated *in vitro* into myocytes that contain myosin heavy chain (MyHC). This demonstrates that a predominant, functional type of stem cell exists in the periventricular region of the adult brain with the intrinsic ability to generate neural and non-neural cells.

As the precise location of NSCs in the adult brain remains controversial^{9–11}, cells were gathered from the two areas with the reported highest NSC content: the ependymal and subventricular zones of the lateral ventricular walls. To identify the putative NSCs within these populations, we examined single-cell suspensions for the expression of a variety of cell surface markers and characteristics with a fluorescence-activated cell sorter (FACS). The frequency of NSCs in sorted sub-populations was determined by the ability of individual cells to generate multipotent neurospheres^{1,12,13}.

Initial NSC enrichment was accomplished by sorting viable cells on the basis of size as assessed by forward light scatter (Fig. 1a). Approximately 80% of the total NSC activity in the unsorted population was found in cells $>12\mu\text{m}$ in diameter (>90 units forward scatter; Table 1). The next enrichment step was achieved by sorting for differential peanut agglutinin (PNA) binding, which has previously been used to fractionate murine haematopoietic stem cells¹⁴. From the FACS profiles, three populations were examined: PNA^{hi} , PNA^{lo} and PNA^{mid} (Fig. 1a). The $>12\mu\text{m}$ PNA^{lo} fraction, representing $1.73 \pm 0.07\%$ (mean $\pm$ s.e.m.; $n = 3$) of the unsorted population, contained the greatest NSC activity with about 1 in 7 cells giving rise to a neurosphere (Table 1). Whereas numerous antigens were detected on the PNA^{lo} population, including CD30, CD34 and CD90.2, heat-stable antigen (HSA, mCD24a) was found to most effectively enrich sub-populations of PNA^{lo} cells with NSCs. The PNA^{lo} HSA^{lo} sub-population (Fig. 1b, c), a discrete population representing $0.27 \pm 0.07\%$ of the unsorted population, comprised NSCs almost exclusively. When cultured under clonal conditions, where one PNA^{lo} HSA^{lo} cell was plated per well, about 80% of the cells (1:1.28) gave rise to neurospheres (Table 1). Furthermore, the PNA^{lo} HSA^{lo} population contained 63.2% of the NSC activity present in the unsorted population, suggesting that the great majority of NSCs in the periventricular region are of this phenotype.

To determine whether all PNA^{lo} HSA^{lo} cells had similar biological potential, clonally derived spheres from PNA^{lo} HSA^{lo} cells (Fig. 2a) were differentiated and assessed by immunocytochemistry for the presence of neurons and glia. All spheres (501 examined) contained astrocytes positive for glial fibrillary acidic protein (GFAP⁺), neurons positive for β -tubulin type III, and oligodendrocytes positive for O4 (Fig. 2b). Indeed, clones derived from PNA^{lo} HSA^{lo} cells possessed all the characteristic *in vitro* properties, such as the ability to self renew (data not shown), of previously described NSCs^{1,6,13}.

To further characterize the PNA^{lo} HSA^{lo} NSCs, we examined the expression of other neural markers. As expected, all PNA^{lo} HSA^{lo} NSCs examined (4,836 cells) expressed nestin, a marker of putative NSCs¹⁵, and were negative for other differentiated cell markers (β -tubulin type III, O4 and GFAP). The lack of GFAP was of interest, as it has been shown that GFAP-expressing cells in the subventricular zone gave rise to neurospheres *in vitro*¹⁰. GFAP expression, however, was determined using an adenoviral vector expressing green fluorescent protein (GFP) under the control of a GFAP promoter, which may be far more sensitive than immunocytochemistry.

PNA^{lo} HSA^{lo} NSCs were found to lack cilia (data not shown) and

Table 1 Flow cytometric purification of NSCs

Cell phenotype	NSC frequency*
Unsorted	1:300
$<7\mu\text{m}$	0
$7-12\mu\text{m}$	1:538
$>12\mu\text{m}$	1:149
$>12\mu\text{m}$ PNA^{hi}	1:1,667
$>12\mu\text{m}$ PNA^{mid}	1:87.9
$>12\mu\text{m}$ PNA^{lo}	1:6.77
$>12\mu\text{m}$ PNA^{lo} $\text{HSA}^{\text{mid-hi}}$	1:15.4 $\pm$ 2.83
$>12\mu\text{m}$ PNA^{lo} HSA^{lo}	1:1.28 $\pm$ 0.11

*NSC frequency is the ratio of the number of neurospheres at 7 d *in vitro* to viable cells plated. Values are the mean of three independent experiments, with standard error of the mean included for final purification steps.

also lacked the expression of the ependymal marker $\text{HSA}^{10,16,17}$, which suggests that the PNA^{lo} HSA^{lo} NSCs reside in the subventricular zone. However, when the ependymal cell layer was labelled with DiI (ref. 11), FACS analysis revealed that $32.1 \pm 4.3\%$ of PNA^{lo} HSA^{lo} cells were DiI positive. Furthermore, the PNA^{lo} HSA^{lo} DiI⁺ population contained equal activity of stem cells to that of the PNA^{lo} HSA^{lo} DiI⁻, whereas other DiI⁺ populations contained extremely low stem cell activity ($<0.001\%$). Thus, the most parsimonious interpretation is that there is a common NSC that resides, or has access to, both the subventricular and ependymal zones, thereby resolving the contentious issue concerning the location of the NSCs⁹⁻¹¹.

Although the *in vitro* data strongly indicate that PNA^{lo} HSA^{lo} is the predominant NSC type, the possibility remains that another type of NSC may account for a considerable proportion of *in vivo* activity. As the primary neuronal output of NSCs in the ventricular region in the adult animal is olfactory neurons¹⁰, we examined the size of the PNA^{lo} HSA^{lo} stem cell population in a mutant strain of mice, *querkopf*, which generates greatly reduced numbers of olfactory neurons in the adult animal and has been shown to have a stem cell deficiency during cortical development⁸. The FACS profile of *querkopf* mice was identical to wild-type littermates and indistinguishable from that in Fig. 1b, except for a sixfold reduction (6.17 ± 1.36) in the percentage of cells in the PNA^{lo} HSA^{lo} compartment. No significant change in the size of any other population was observed, and the stem cell activity was still restricted to the PNA^{lo} HSA^{lo} compartment in the mutant. Thus, the only detectable change in the cellular content of the periventricular region associated with this loss of neuronal production is a specific depletion of the PNA^{lo} HSA^{lo} NSCs, strongly supporting a central functional role for these cells *in situ*.

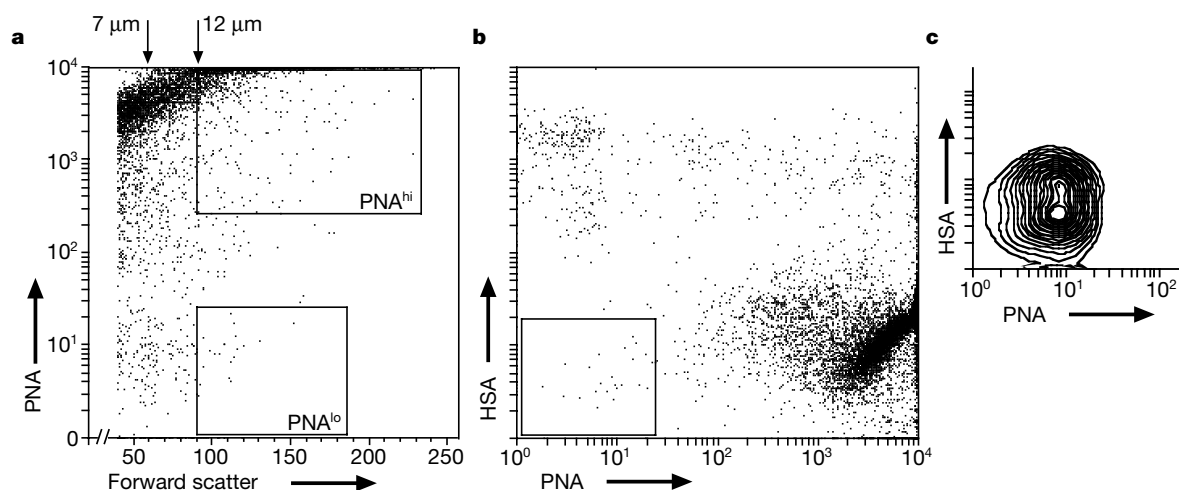


Figure 1 Flow cytometric purification of NSCs. **a**, Viable CNS cells sorted on forward light scatter (<7 , $7-12$, $>12\mu\text{m}$) and PNA expression (PNA^{lo} , PNA^{mid} and PNA^{hi}); NSC activity was predominantly found in the $>12\mu\text{m}$ PNA^{lo} cells (boxed). **b**, Cells $>12\mu\text{m}$ were next

sorted on the basis of HSA immunoreactivity; NSC activity was predominantly found in the PNA^{lo} HSA^{lo} cells (boxed). **c**, Contour plot of boxed region in **b** shows antigenic expression of this population in more detail.

To further examine their *in vivo* potential, freshly isolated PNA^{lo} HSA^{lo} NSCs were injected into the lateral ventricles of E12.5 mice. Donor PNA^{lo} HSA^{lo} NSCs were isolated from BU5X transgenic mice, which express both *lacZ*¹⁸ and enhanced GFP¹⁹ but show identical stem cell characteristics to those of CBA mice (data not shown). When examined 4 weeks after birth, GFP⁺ neurons and glia were detected in the cortex (Fig. 2g, h), demonstrating that the purified populations of stem cells retain the capacity to differentiate *in vivo* and do not require accessory cell stimulation.

Recently, the possibility has been raised that bone marrow cells have the potential to give rise to neurons *in vivo*^{20,21}. FACS analysis of PNA^{lo} HSA^{lo} cells, however, revealed an absence of haematopoietic cell surface antigens CD34, CD90.2, CD117 and CD135 and the

endothelial marker CD31. Furthermore, most PNA^{lo} HSA^{lo} cells were brightly labelled with the DNA-binding dye Hoechst 33342 (data not shown), whereas activity of haematopoietic stem cells is unambiguously found in the Hoechst 'low'-staining compartment²². Taken together, these data strongly suggest that the PNA^{lo} HSA^{lo} NSC is distinct from the bone-marrow-derived stem cell and that the latter is unlikely to be important in neuronal production in the adult CNS.

To examine whether NSCs had the innate potential to give rise to non-neural cell types, we co-cultured freshly isolated PNA^{lo} HSA^{lo} GFP⁺ NSCs with C2C12 cells under conditions previously described to give rise to myogenic cells²³. Many GFP-expressing cells assumed a spindle-like, myocyte-like shape after 4 d *in vitro* (Fig. 2c), often containing multiple nuclei, which is characteristic of myotubes (arrow, Fig. 2d). Importantly, GFP was co-expressed in cells expressing the myogenic markers α -actinin-2 (Fig. 2e) and fast MyHC (Fig. 2f): in the case of myotubes, often only one bright GFP⁺ nucleus among several GFP⁻ nuclei; in the case of myocytes, a single GFP⁺ nucleus (arrow, Fig. 2e; Fig. 2f). Other populations of sorted cells did not exhibit myogenic potential (data not shown). As clonal analysis and direct observation revealed no evidence of significant cell division occurring after plating (data not shown), the frequency of NSCs with myogenic potential could be determined directly. We found that $57 \pm 5.67\%$ of the plated PNA^{lo} HSA^{lo} GFP⁺ NSCs differentiated into spindle-shaped MyHC⁺ myocytes, or myotubes. This demonstrates that the potential for muscle differentiation is intrinsic to most PNA^{lo} HSA^{lo} NSCs, and that the same cells have the potential to generate neural and muscle cells. This is an unequivocal demonstration of pluripotential activity in a freshly isolated stem cell.

Overall, the data strongly indicate that the predominant type of NSC in the subventricular zone of the forebrain has the phenotype PNA^{lo} HSA^{lo} nestin⁺, and the intrinsic potential to differentiate into both neural and non-neural cells. The identification and purification of NSCs provide a means to study directly the regulation of NSC differentiation both *in vitro* and *in vivo*, with the ultimate aim of stimulating endogenous neuronal production to replace damaged neuronal tissue. These methods permit the NSCs to be compared directly to stem cells from other tissues, including embryonic stem cells, to establish whether they have similar biological and molecular properties. □

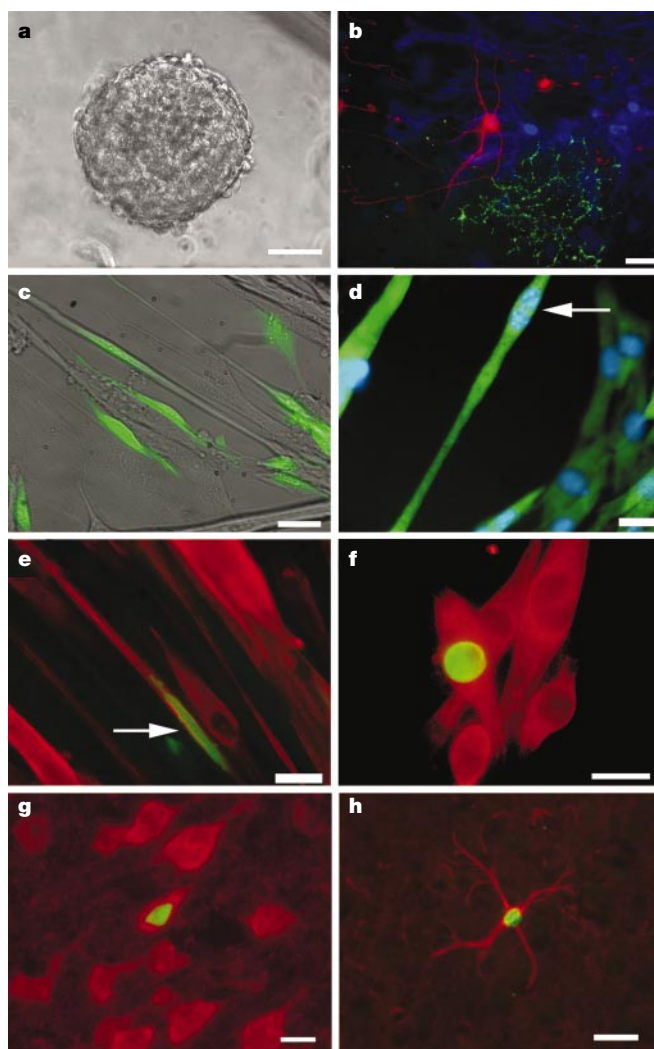


Figure 2 PNA^{lo} HSA^{lo} adult NSCs give rise to neural and muscle cell types. A single PNA^{lo} HSA^{lo} cell generates a neurosphere (a) that, when transferred to differentiating conditions, has progeny (b) including neurons (red, β -tubulin type III), astrocytes (blue, GFAP) and oligodendrocytes (green, O4). When co-cultured with C2C12 cells, freshly isolated PNA^{lo} HSA^{lo} GFP⁺ NSCs assume a spindle-like morphology (c) often with multiple nuclei (DAPI⁺, arrow, d), suggestive of muscle cells. Myogenic conversion of NSCs confirmed by anti-fast MyHC type II (red, e) and anti- α -actinin-2 (red, f) immunoreactive cells containing GFP⁺ nuclei. In many cases, immunostaining procedures reduced the intensity of cytoplasmic GFP to below detectable levels, thus GFP expression was found predominantly in the nucleus. When injected into C57BL/6J pups *in utero*, donor-derived (β -gal⁺, green) PNA^{lo} HSA^{lo} NSCs generated NeuN⁺ (red, g) neurons and GFAP⁺ (red, h) astrocytes. Scale bars: a, 80 μ m; b, c, e, 40 μ m; d, g, h, 20 μ m; f, 15 μ m.

Methods

Tissue preparation and FACS analysis

Adult CBA mice (8–10 weeks old) were killed by cervical dislocation, and periventricular tissue was dissected (essentially as described²⁴) into HEPES-buffered Eagle's medium (HEM). CNS tissue was diced then transferred to Mg²⁺/Ca²⁺-free HBSS (containing 10 mM HEPES at pH 7.6, 200 g ml⁻¹ EDTA, 0.5 mM trypsin, 0.001% DNase), for 20 min at 37 °C. Six millilitres of HEM with 5% fetal calf serum (FCS) was added and tissue collected by centrifugation (7 min at 100g). The supernatant was removed and the pellet triturated in 200 μ l PBS (pH 7.4), producing a single-cell suspension that was subsequently passed through a 70- μ m cell strainer (Falcon) to remove debris. For immunostaining, this suspension was incubated for 20 min at 4 °C with FITC-conjugated PNA (1:200; Vecta), or PNA-FITC combined with PE-conjugated mCD24a (1:200; clone M1/69, PharMingen), then rinsed twice with PBS by centrifugation, and finally with PBS, 1% FCS, propidium iodide (100 μ g ml⁻¹; Molecular Probes) to label dead cells before sorting on a FACS II (Becton-Dickinson) flow cytometer. Cell viability was typically >95% and all FACS gates were set using unlabelled cells. Antibodies against CD31, 34, 90.2, 117 and 135 (1:200; PharMingen) were directly conjugated to either FITC or PE where appropriate.

Cell culture

Neurospheres were generated as previously described¹³. Briefly, sorted cells were plated with bFGF (bovine recombinant, 10 ng ml⁻¹; Boehringer) and EGF (receptor grade, 20 ng ml⁻¹; Collaborative Research) in NS-A basal serum free media (basal media; Euroclone) containing 2 mM L-glutamine, 0.6% glucose, 60 μ M putrescine, 20 nM progesterone, 30 nM sodium selenite, 25 μ g ml⁻¹ insulin and 100 μ g ml⁻¹ apo-transferrin (Sigma). Initially, cells were seeded at 10 viable cells cm⁻², final enrichment steps at one cell per well. The ratio of number of spheres formed after 7 d *in vitro* to number of cells plated is the NSC frequency. Subsequent passaging of primary spheres was performed by

mechanically dissociating collected spheres (centrifuged for 10 min at 100g) into a single-cell suspension and replating in basal media containing EGF and FGF-2 (complete medium) as described^{1,12}. Spheres were differentiated by transfer to glass coverslips coated with poly-L-ornithine (one sphere per coverslip) in complete medium for 1 d, basal medium for 1 d, then basal medium with 1% FCS for 4–5 d, then assessed by immunocytochemistry for neurons and glia.

Freshly isolated NSCs from BU5X mice were co-cultured with C2C12 myogenic cells (5×10^3 cells cm^{-2}) for 2 d in DME medium (Gibco) containing 20% heat-inactivated FCS, then in DME supplemented with 1% normal horse serum (CSL) for an additional 2–4 d *in vitro*. Cultures were fixed for 5 min with 4% paraformaldehyde then processed for immunocytochemistry.

Immunocytochemistry

Double-antigen immunocytochemistry on neurospheres was performed as described previously^{11,12}, using monoclonal antibody to β -tubulin type III (Sigma), GFAP antisera (Dako) and monoclonal antibody to O4 (immunoglobulin- μ , IgM, Boehringer). Neural-derived muscle cells were identified by simultaneous detection of endogenous GFP and muscle cell types identified by mouse monoclonal antibodies to fast MyHC (NCL-MHCF, 1:10, Novocastra) or α -actinin-2 (1:750; ref. 25). These antigens were detected by appropriate TRITC-conjugated IgG secondary antibody (1:200; Southern Biotech). The proportion of neural cells that differentiated into muscle cells was determined by counting the number of GFP⁺ nuclei expressing myogenic markers divided by the number of NSCs plated (counted 6 h after plating). To detect the progeny of freshly isolated BU5X NSCs injected into (C57BL/6 $\times$ DBA/2) F₁ hybrid pups (E12.5), mice were perfused with 4% paraformaldehyde, their brain sectioned (10- μ m slices) and immunocytochemistry performed to double-label, donor-derived cells immunoreactive to β -gal (anti- β -gal, Chemicon) with known markers for astrocytes (anti-GFAP, Chemicon) or neurons (anti-NeuN, Chemicon). Images were captured on a Nikon Diaphot microscope using a KX-85 (Apogee) camera.

Received 21 March; accepted 29 June 2001.

- Reynolds, B. A. & Weiss, S. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* **255**, 1707–1710 (1992).
- Richards, L. J., Kilpatrick, T. J. & Bartlett, P. F. De novo generation of neuronal cells from the adult mouse brain. *Proc. Natl Acad. Sci. USA* **89**, 8591–8595 (1992).
- McKay, R. Stem cells in the central nervous system. *Science* **276**, 66–71 (1997).
- Gage, F. H. Mammalian neural stem cells. *Science* **287**, 1433–1438 (2000).
- Clarke, D. L. *et al.* Generalized potential of adult neural stem cells. *Science* **288**, 1660–1663 (2000).
- Bjornson, C., Rietze, R., Reynolds, B., Magli, M. & Vescovi, A. Turning brain into blood: a hematopoietic fate adopted by adult neural stem cells in vivo. *Science* **283**, 534–537 (1999).
- Bartlett, P. F. Pluripotential hemopoietic stem cells in adult mouse brain. *Proc. Natl Acad. Sci. USA* **79**, 2722–2725 (1982).
- Thomas, T., Voss, A. K., Chowdhury, K. & Gruss, P. Querkopf, a MYST family histone acetyltransferase, is required for normal cerebral cortex development. *Development* **127**, 2537–2548 (2000).
- Chiasso, B. J., Tropepe, V., Morshead, C. M. & van der Kooy, D. Adult mammalian forebrain ependymal and subependymal cells demonstrate proliferative potential, but only subependymal cells have neural stem cell characteristics. *J. Neurosci.* **19**, 4462–4471 (1999).
- Doetsch, F., Caille, L., Lim, D. A., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. Subventricular zone astrocytes are neural stem cells in the adult mammalian brain. *Cell* **97**, 703–716 (1999).
- Johansson, C. B. *et al.* Identification of a neural stem cell in the adult mammalian central nervous system. *Cell* **96**, 25–34 (1999).
- Gritti, A. *et al.* Multipotent neural stem cells from the adult mouse brain proliferate and self-renew in response to basic fibroblast growth factor. *J. Neurosci.* **16**, 1091–1100 (1996).
- Gritti, A. *et al.* Epidermal and fibroblast growth factors behave as mitogenic regulators for a single multipotent stem cell-like population from the subventricular region of the adult mouse forebrain. *J. Neurosci.* **19**, 3287–3297 (1999).
- Salner, A. L., Obbagy, J. E. & Hellman, S. Differing stem cell self-renewal of lectin-separated murine bone marrow fractions. *J. Natl Cancer Inst.* **68**, 639–641 (1982).
- Lendahl, U., Zimmerman, L. B. & McKay, R. D. CNS stem cells express a new class of intermediate filament protein. *Cell* **60**, 585–595 (1990).
- Calao, V., Chazal, G., Nielsen, P. J., Rougon, G. & Moreau, H. mCD24 expression in the developing mouse brain and in zones of secondary neurogenesis in the adult. *Neuroscience* **73**, 581–594 (1996).
- Shewan, D. *et al.* mCD24, a glycoprotein transiently expressed by neurons, is an inhibitor of neurite outgrowth. *J. Neurosci.* **16**, 2624–2634 (1996).
- Tam, P. P. & Tan, S. S. The somitogenic potential of cells in the primitive streak and the tail bud of the organogenesis-stage mouse embryo. *Development* **115**, 703–715 (1992).
- Hadjantonakis, A. K., Gertsenstein, M., Ikawa, M., Okabe, M. & Nagy, A. Generating green fluorescent mice by germline transmission of green fluorescent ES cells. *Mech. Dev.* **76**, 79–90 (1998).
- Brazelton, T. R., Rossi, F. M. V., Keshet, G. I. & Blau, H. M. From marrow to brain: expression of neuronal phenotypes in adult mice. *Science* **290**, 1775–1779 (2000).
- Mezey, E., Chandross, K. J., Harta, G., Maki, R. A. & McKeacher, S. R. Turning blood into brain: cells bearing neuronal antigens generated in vivo from bone marrow. *Science* **290**, 1779–1782 (2000).
- Goodell, M. A. *et al.* Dye efflux studies suggest that hematopoietic stem cells expressing low or undetectable levels of CD34 antigen exist in multiple species. *Nature Med.* **3**, 1337–1345 (1997).
- Galli, R. *et al.* Skeletal myogenic potential of human and mouse neural stem cells. *Nature Neurosci.* **3**, 986–991 (2000).
- Lois, C. & Alvarez-Buylla, A. Proliferating subventricular zone cells in the adult mammalian forebrain can differentiate into neurons and glia. *Proc. Natl Acad. Sci. USA* **90**, 2074–2077 (1993).
- North, K. N. & Beggs, A. H. Deficiency of a skeletal muscle isoform of α -actinin (α -actinin-3) in merosin-positive congenital muscular dystrophy. *Neuromuscul. Disord.* **6**, 229–235 (1996).

Acknowledgements

This work was supported by grants from the NH-MRC of Australia, the Australasian Spinal Research Trust, and the Motorneuron Disease Research Institute. We express our gratitude to P. Tam for supplying BU5X mice; S. S. Tan for transplantation assistance; E. Hardeman for supply and assistance with C2C12 myocytes and muscle-specific antibodies and J. Coonan for assistance with microinjections.

Correspondence and requests for materials should be addressed to P.F.B. (e-mail: bartlett@wehi.edu.au).

Antibodies inhibit prion propagation and clear cell cultures of prion infectivity

David Peretz^{*}, R. Anthony Williamson[†], Kiyoshi Kaneko^{*‡}, Julie Vergara^{*}, Estelle Leclerc[†], Gerold Schmitt-Ulms^{*}, Ingrid R. Mehlhorn^{*}, Giuseppe Legname^{*}, Mark R. Wormald[§], Pauline M. Rudd[§], Raymond A. Dwek[§], Dennis R. Burton^{†||} & Stanley B. Prusiner^{*¶}

^{*} Institute for Neurodegenerative Diseases and Departments of [†] Neurology and [‡] Biochemistry and Biophysics, University of California, San Francisco, California 94143-0518, USA

[†] Departments of Immunology and ^{||} Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, USA

[§] The Glycobiology Institute, Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK

Prions are the transmissible pathogenic agents responsible for diseases such as scrapie and bovine spongiform encephalopathy. In the favoured model of prion replication, direct interaction between the pathogenic prion protein (PrP^{Sc}) template and endogenous cellular prion protein (PrP^C) is proposed to drive the formation of nascent infectious prions^{1,2}. Reagents specifically binding either prion-protein conformer may interrupt prion production by inhibiting this interaction. We examined the ability of several recombinant antibody antigen-binding fragments (Fabs) to inhibit prion propagation in cultured mouse neuroblastoma cells (ScN2a) infected with PrP^{Sc}. Here we show that antibodies binding cell-surface PrP^C inhibit PrP^{Sc} formation in a dose-dependent manner. In cells treated with the most potent antibody, Fab D18, prion replication is abolished and pre-existing PrP^{Sc} is rapidly cleared, suggesting that this antibody may cure established infection. The potent activity of Fab D18 is associated with its ability to better recognize the total population of PrP^C molecules on the cell surface, and with the location of its epitope on PrP^C. Our observations support the use of antibodies in the prevention and treatment of prion diseases and identify a region of PrP^C for drug targeting.

To study inhibition of prion propagation by antibodies, we used recombinant prion protein-specific Fabs D13, D18, R1, R2, E123, E149 and R72 (refs 3–6). The binding epitopes and affinity constants of the antibodies for recombinant prion protein are shown in Supplementary Information Table 1. Fab R72 does not recognize prion protein in surface plasmon resonance (SPR) or on the cell surface, but does bind to PrP^C coated onto the surface of wells for enzyme-linked immunosorbent assay (ELISA)⁴.

A range of concentrations of each antibody was added to ScN2a cultures for 7 days. Cells were then collected and the level of PrP^{Sc} in the culture analysed by immunoblotting. The level of PrP^{Sc} in cells

[¶] Present address: National Institute of Neuroscience, Tokyo 187-8502, Japan.

treated with Fabs D13, D18, R1 and R2, compared with that in non-treated cells, was dramatically reduced in a dose-dependent manner (Fig. 1). By this analysis, Fabs D13 and D18 seem to be about equally effective: their values for 50%-inhibitory concentration (IC_{50}) were $0.6 \mu g ml^{-1}$ (12 nM) and $0.45 \mu g ml^{-1}$ (9 nM), respectively. Fabs R1 and R2 were slightly less efficient, with IC_{50} values of $2.5 \mu g ml^{-1}$ (50 nM) and $2.0 \mu g ml^{-1}$ (40 nM), respectively. In contrast, 7-day treatment with Fabs E123, E149 or R72, or 5-day treatment with polyclonal immunoglobulin- γ (IgG) recognizing both transmembrane and glycosyl phosphatidylinositol (GPI)-anchored forms of mouse neural cell-adhesion molecules (N-CAM)⁷, did not reduce the level of PrP^{Sc} in ScN2a cultures even when these antibodies were used at high concentrations (Fig. 1). During these experiments, the levels of PrP^C and glyceraldehyde-3-phosphate dehydrogenase in antibody-treated and untreated cells were found to be invariant (see Supplementary Information Fig. 1), indicating that the prion protein-specific antibodies used produced no cytotoxic effects that may have indirectly compromised the production of PrP^{Sc}.

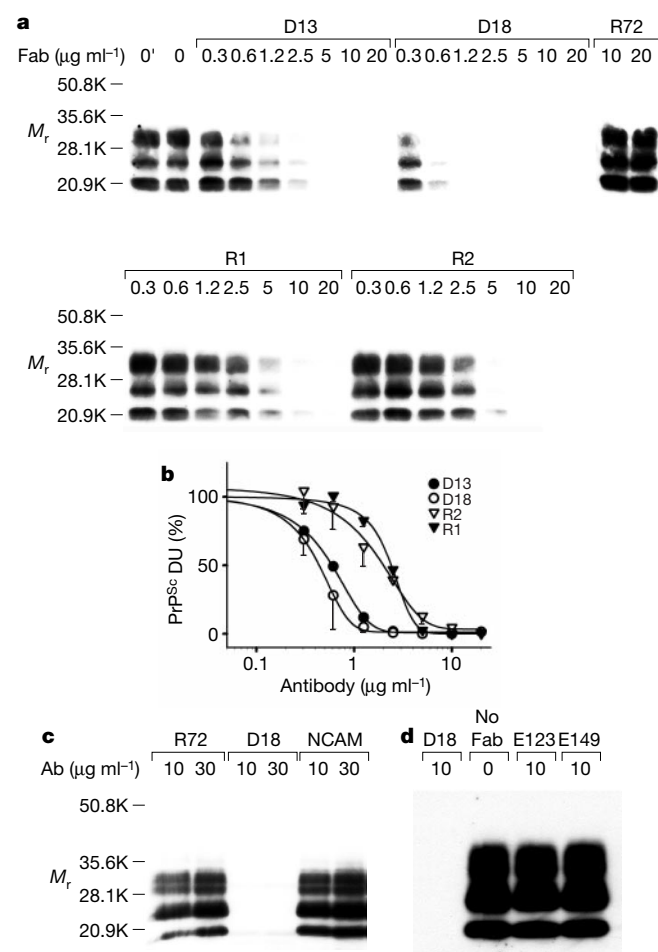


Figure 1 Dose-dependent inhibition of PrP^{Sc} formation in cells by prion protein-specific recombinant antibody Fabs. **a**, PrP^{Sc} levels in ScN2a cells were measured by immunoblotting after 7 d of culture in the presence of antibodies D13, D18, R72, R1 or R2 at 0–20 $\mu g ml^{-1}$. Lane 0' indicates the level of PrP^{Sc} in the ScN2a culture before antibody treatment. Protein size is shown as relative molecular mass (M_r), in thousands (K). **b**, Densitometric measurement of PrP^{Sc} bands identified in the immunoblot in **a**. Values are given as densitometric units (DU), where 100% is equivalent to the intensity of the PrP^{Sc} band in the absence of antibody treatment and 0% denotes undetectable levels of PrP^{Sc} in the culture (no band). Data represent the mean from three independent experiments. **c**, **d**, PrP^{Sc} levels in ScN2a cells cultured for 5 d in the presence of anti-N-CAM polyclonal IgG (**c**), or for 7 d in the presence of Fabs E123 and E149 (**d**).

We then determined whether PrP^{Sc} remained undetectable after removal of prion protein-specific antibody. ScN2a cells were independently passaged (serially subcultured) for at least 7 days in the presence of $10 \mu g ml^{-1}$ of each of the recombinant Fabs. Antibody was then removed from the culture and the cells passaged for an additional period in Fab-free medium, after which time the level of PrP^{Sc} was re-measured (Fig. 2). PrP^{Sc} concentrations in ScN2a cells passaged for 1 week in the presence of Fab D18 were reduced to non-detectable levels, but returned to about 50% of the level of an untreated control culture after 1 additional week of growth in the absence of D18. However, if cells were cultured for a 2-week period in the presence of Fab D18 then PrP^{Sc} remained at undetectable levels after 4 additional weeks of culture in antibody-free medium. Similarly, when prion-infected cells were treated with Fab D13 for 3 consecutive weeks followed by 1 week of growth in media without antibody, no PrP^{Sc} could be detected, although after an additional week in culture without Fab, PrP^{Sc} increased back to 5% of the level found in untreated control culture. If, however, cells were subjected to 4 consecutive weeks of treatment with Fab D13, PrP^{Sc} remained below the level of detection after 4 weeks of culture without antibody (data not shown). When used at a concentration of $10 \mu g ml^{-1}$ neither Fab R1 nor Fab R2 was sufficiently potent to prevent the re-emergence of PrP^{Sc} in the culture after antibody was removed. Fab R72 had no impact on the level of PrP^{Sc} in the culture after 3 or 9 (data not shown) consecutive weeks of treatment.

As a second measure of prion titre, we performed bioassays in which CD-1 Swiss mice were inoculated with antibody-treated ($10 \mu g ml^{-1}$) and untreated ScN2a cells. Mice inoculated intracerebrally with Fab D18-, D13- or even R2-treated cells were free of disease after 265 days, whereas mice inoculated with untreated or R72-treated cells had a mean incubation time to disease of 169 and 165 days, respectively. The prolonged incubation times correspond to a reduction of over three orders of magnitude in the infectious prion titre in treated cells⁸.

The above data might be taken to indicate that PrP^{Sc} levels rapidly diminish in the presence of specific antibody. Indeed, other studies describing the inhibition of prion propagation^{9–13} have been interpreted in this way. However, if an antibody or other reagent curtails the formation of nascent PrP^{Sc} molecules, each round of cell division may serve to dilute the effective concentration, but not necessarily the total amount, of PrP^{Sc} within the culture.

To more thoroughly analyse the kinetics of prion clearance,

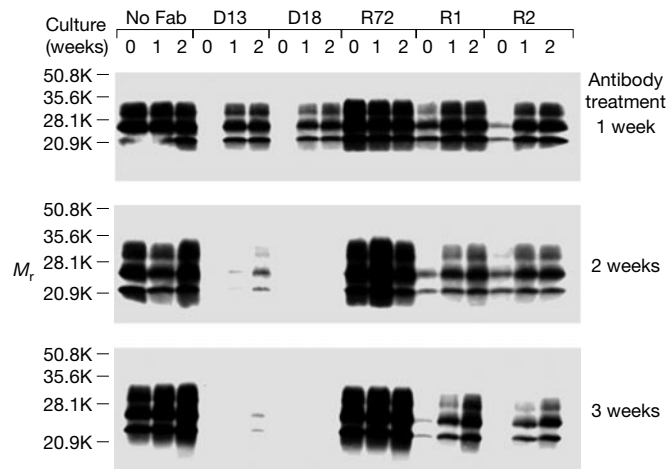


Figure 2 Elimination of PrP^{Sc} from prion-infected cells treated with antibodies. ScN2a cells were cultured for 1, 2 or 3 weeks in $10 \mu g ml^{-1}$ of either Fab D13, D18, R72, R1 or R2, or in the absence of antibody. The level of PrP^{Sc} in the culture was then analysed by immunoblot either immediately after the termination of antibody treatment (lane 0) or after cells were passaged for an additional 1 (lane 1) or 2 (lane 2) weeks in the absence of antibody.

ScN2a cells were independently grown in the presence of $10 \mu\text{g ml}^{-1}$ of prion protein-specific Fabs. Cells were collected after 1, 2, 3 or 4 days of antibody treatment, and the total mass of cell protein was determined in each case as a measure of cell number. The PrP^{Sc} concentrations were determined by immunoblotting and the total PrP^{Sc} in the culture at each point was calculated by factoring in the total cell mass in each case (Fig. 3).

In agreement with earlier results, Fab D18 was found to be the most effective antibody. The time taken from the initial treatment with Fab D18 to eliminate 50% of PrP^{Sc} from the cells (half-life) was 28 h. The half-life of PrP^{Sc} in ScN2a cells is thought to exceed 24 h (ref. 14), suggesting that, at a concentration of $10 \mu\text{g ml}^{-1}$, Fab D18 is able to completely abolish prion propagation and that pre-existing PrP^{Sc} is subsequently eliminated from the cells. This finding indicates that a certain amount of PrP^{Sc} is continuously expunged from ScN2a cultures through normal degradation pathways. Fab D13 was the next most potent antibody, also lowering the level of PrP^{Sc} in the culture, but to a lesser extent than Fab D18, indicating that there may be a minimal level of residual PrP^{Sc} synthesis in the presence of this Fab. Fabs R1 and R2, although reducing the rate of prion propagation in ScN2a cells, are not sufficiently effective to yield a reduction in the quantity of PrP^{Sc} in the culture. In untreated cultures, or cultures treated with Fab R72, prion propagation remained unaffected, and PrP^{Sc} levels increased in tandem with growth in the ScN2a cell population.

The inhibitory effect we observe is most readily explained by antibodies binding specifically to PrP^C molecules on the cell surface and thereby hindering the docking of PrP^{Sc} template or a cofactor critical for the conversion of PrP^C to PrP^{Sc}. In agreement with this hypothesis, Fab D18, which was by far the most effective antibody we evaluated, was distinguished by its capacity to bind a signifi-

cantly greater number of cell-surface PrP^C molecules than any of the other antibodies tested (see Supplementary Information Fig. 2).

To address whether the region of prion protein bound by each antibody was of intrinsic importance to its inhibitory potency, we compared inhibition at conditions in which equal amounts of two different Fabs were bound to the ScN2a cell surface. When used at concentrations of $0.6 \mu\text{g ml}^{-1}$, equivalent amounts of Fabs D18 and D13 were bound to the cell surface but D18 inhibited prion replication much more efficiently (Fig. 1a). Similarly, at concentrations of $0.6 \mu\text{g ml}^{-1}$ and $2.5 \mu\text{g ml}^{-1}$, respectively, Fabs D18 and R1 bound equivalently to ScN2a cells, but D18 was clearly more effective in reducing the level of PrP^{Sc} in the culture. Finally, at a concentration of $2.5 \mu\text{g ml}^{-1}$, Fabs D13 and R1 bound equivalently to the cell surface, but D13 more actively reduced PrP^{Sc} synthesis.

The demonstration that the region of PrP^C bound by a given antibody is a critical determinant of its inhibitory capacity provides insight into the specific machinery of prion propagation. The known binding epitope of Fab D18 spans prion protein residues 132–156 (ref. 5) and incorporates helix A (residues 145–155) of PrP^C (Fig. 4). Spatially, this sequence is positioned on the opposite face of the protein from residues Q167, Q171, T214 and Q218, which are thought to participate in binding an auxiliary molecule essential to prion propagation^{15,16}. We therefore argue that D18 operates mechanistically by directly blocking or modifying PrP^C interaction with PrP^{Sc}, rather than by inhibiting the binding of a cofactor. A number of other reports also identify the 132–156 region of the protein as critical to prion synthesis and between-species transmission^{17–20}, although prion propagation can proceed in the absence of sequence between residues 140 and 175 (ref. 21), indicating that helix A of PrP^C is not a prerequisite for PrP^{Sc} binding. Together, these studies identify the 132–140 portion of PrP^C as a logical target for development of anti-prion drugs.

Significantly, with the exception of Fabs R1 and R2, none of the antibodies we have used in this study competes efficiently with the others for binding to cell-surface PrP^C (E.L., unpublished data). This suggests that, in therapeutic strategies, the antibodies may be used in combination to achieve maximum inhibitory effect.

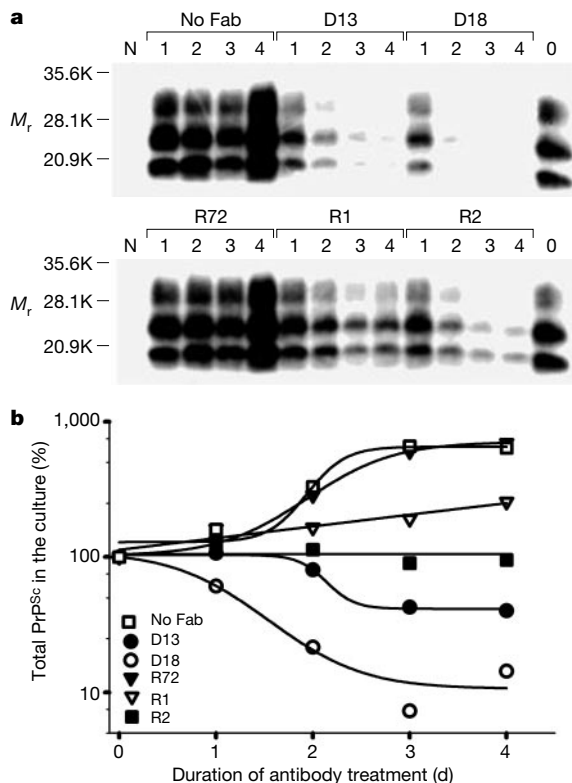


Figure 3 Time course of antibody-mediated PrP^{Sc} clearance. **a**, The level of PrP^{Sc} in ScN2a cells grown for 1, 2, 3 or 4 d in the presence of prion protein-specific Fabs ($10 \mu\text{g ml}^{-1}$) was determined by immunoblotting. **b**, The effect of antibody treatment on the total amount of PrP^{Sc} in ScN2a cell cultures. Lane N indicates N2a cells treated with proteinase K. The data represent the mean of three experiments.

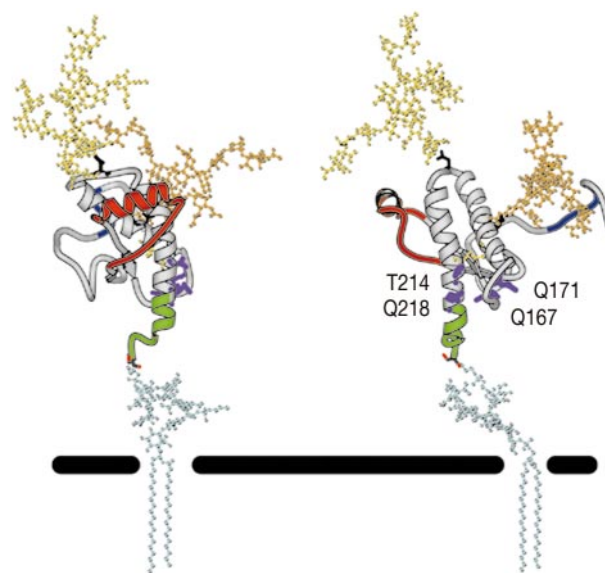


Figure 4 Regions of sequence recognized by Fabs D13 (blue), D18 (red), and R1 and R2 (green) superimposed onto two views of the structure of recombinant prion protein (90–231) determined by nuclear magnetic resonance²⁷. Carbohydrate moieties linked to Asn 180 and Asn 196 are shown in orange and yellow, respectively^{28,29}. The COOH-terminal GPI anchor is shown in cyan extending into the cell membrane (black). Side chains of residues Q167, Q171, T214 and Q218, which are proposed to bind to a cellular cofactor critical to prion propagation, are included (purple).

We have shown that antibodies binding to defined regions of PrP^C effectively inhibit prion replication. For *in vivo* applications, Fab fragments have the disadvantage of a short half-life, and may not efficiently traverse from the peripheral circulation into the central nervous system. Whole-antibody molecules prepared from the Fabs will probably be more useful, but may require engineering to prevent the recruitment of immunological effector mechanisms to antibody-coated cells^{22,23}. Our studies, in combination with advances targeting Alzheimer disease^{24,25}, indicate that specific antibodies may become a powerful weapon in the fight against neurodegenerative diseases associated with the accumulation of misfolded proteins. □

Methods

Expression and purification of antibodies

Escherichia coli strain 33B6 was transformed with plasmids encoding prion protein-specific Fabs, and fermented for 48 h using a Biostat B controller (Braun) and 11 media containing MT-8 salts (0.26 g l⁻¹ K₂HPO₄, 0.13 g l⁻¹ NaH₂PO₄·2H₂O, 0.5 g l⁻¹ NH₄SO₄, 0.1 g l⁻¹ C₆H₅Na₃O₇·2H₂O and 0.15 g l⁻¹ KCl), 0.5 g isoleucine, 20% NZ amines, 20% yeast extract, 1 mM MgSO₄, 50% glucose, trace metals and 100 µg ml⁻¹ ampicillin. Bacterial paste was resuspended in 5 volumes of 2 mM imidazole, 20 mM Na₂HPO₄ and 250 mM NaCl, pH 7.0, and processed twice in a Microfluidizer M-110 EH (Microfluidics). The processed paste was titrated to 0.1% polyethylenimine (5% stock solution, pH 8.0) and stirred at 4 °C for 30 min, then centrifuged at 12,000g for 30 min at 4 °C. The supernatant was diluted in an equal volume of 20 mM imidazole, pH 7.0, and loaded onto an SP Fast Flow Sepharose column (Amersham). Recombinant Fab was eluted with a linear gradient of 0–100% of 20 mM imidazole, 500 mM sodium acetate, pH 7.0, then directly applied to a column for immobilizing by metal-affinity chromatography (IMAC). Antibody was eluted from this column with 200 mM imidazole, pH 7.0, then thoroughly dialysed at 4 °C against 10 mM Tris-HCl buffer, pH 7.2. The dialysed samples were further purified by elution from a QS Fast Flow Sepharose column using a linear gradient of 0–100% of 10 mM Tris-HCl buffer, pH 7.2, and 500 mM NaCl, and sterilized by filtration.

Cell culture

Mouse neuroblastoma cells (N2a) were obtained from the American Type Culture Collection. Prion-infected mouse neuroblastoma (ScN2a) cells have been described previously^{8,26}. Stock cultures of ScN2a and N2a cells were maintained in MEM, 10% fetal bovine serum (FBS), 2 mM Glutamax (GIBCO BRL), 100 units ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin sulphate in a humidified 37 °C incubator with 5% CO₂. Cells were split 1:15 weekly using 0.05% (w/v) trypsin-EDTA (GIBCO BRL).

Studies of antibody inhibition

Antibodies were added to 2 × 10⁵ ScN2a cells and incubated for an appropriate time according to individual experimental protocols. The cells were fed three times per week with replacement media containing the appropriate amount of antibody. In inhibition experiments in which ScN2a cultures required splitting, cells were detached from culture plates using cell dissociation buffer (GIBCO BRL), rather than with trypsin, because enzymatic activity may have modulated the level of PrP^{Sc}. Cells were collected *in situ* by washing three times with calcium- and magnesium-free PBS and resuspending in 1 ml lysis buffer (10 mM Tris, pH 7.5, containing 150 mM NaCl, 0.5% sodium deoxycholate, 0.5% nonident P-40). Cell nuclei were removed from the lysate by centrifugation at 2,000g for 2 min, and the protein concentration of the supernatant measured by bicinchoninic acid assay (BCA, Pierce).

Quantification of PrP^{Sc}

ScN2a cell lysates (1 mg ml⁻¹) were treated with 20 µg ml⁻¹ of proteinase K (total protein: enzyme ratio, 50:1) for 1 h at 37 °C. Proteolytic digestion was terminated by the addition of phenylmethyl sulphonyl fluoride (PMSF) to a final concentration of 2 mM. Forty micrograms of the digested cell lysate was mixed with an equal volume of 2 × non-reducing SDS sample buffer, boiled for 5 min, cleared by centrifugation and resolved by SDS-PAGE (14%). Samples were electroblotted onto membranes of polyvinylidene fluoride (PVDF) and blocked with 5% (w/v) non-fat milk protein in calcium- and magnesium-free PBS. Prion protein was detected using 0.5 µg ml⁻¹ D18 antibody cross-linked to pre-activated amine-reactive horseradish peroxidase (Pierce). Blots were developed with enhanced chemiluminescence (ECL) reagent (Amersham) for 1 min, and exposed to ECL Hypermax film (Amersham). Densitometric scanning of PrP^{Sc} bands was performed with a Chemi Imager 4000 Low Light Imaging System using AlphaEase software version 3.3e (Alpha Innotech). Apparent amounts of PrP^{Sc} (densitometric units) were plotted as a percentage of PrP^{Sc} found in equivalent untreated ScN2a cell cultures on the day of collection. Levels of PrP^C were determined similarly, except that N2a lysates were not digested with proteinase K.

Bioassays

Antibody-treated (60 d) and untreated ScN2a cells were resuspended in 1 ml of PBS. Thirty microlitres of the cell resuspension were inoculated intracerebrally into groups of

10 CD-1 Swiss mice. Mice were scored daily for early and late onset of clinical signs of scrapie. Infectious titres of mouse prions in CD-1 Swiss mice were derived by use of the equation

$$\log(T) = 1.52 + \log(D) + (185 - Y)/12.66$$

where T is the ID₅₀ (in units ml⁻¹), D is the dilution (defined as the fractional dilution of the diluted sample), and Y is the mean interval from inoculation to onset of terminal illness in days⁸.

Flow cytometry

ScN2a cells were resuspended with cell-dissociation buffer (GIBCO BRL) and washed twice with FACS buffer (MEM, 5% (v/v) cell-dissociation buffer and 2% (v/v) FBS). Increasing concentrations of prion protein-specific Fabs were added to aliquots of 10⁶ cells. After incubation for 15 min at room temperature, cells were washed four times with FACS buffer and stained for 15 min at room temperature with a 1:200 dilution in FACS buffer of fluorescein isothiocyanate (FITC)-conjugated goat anti-human IgG Fab (Jackson Immunologicals). Cells were then washed three times in FACS buffer, fixed with freshly prepared 0.4% (v/v) paraformaldehyde and analysed using the Becton Dickinson FACScan instrument.

Nomenclature

Numbering of residues corresponds to mouse prion protein throughout.

Received 19 March; accepted 23 July 2001.

- Telling, G. C. *et al.* Evidence for the conformation of the pathologic isoform of the prion protein enciphering and propagating prion diversity. *Science* **274**, 2079–2082 (1996).
- Prusiner, S. B. Prions. *Proc. Natl Acad. Sci. USA* **95**, 13363–13383 (1998).
- Williamson, R. A. *et al.* Circumventing tolerance in order to generate autologous monoclonal antibodies to the prion protein. *Proc. Natl Acad. Sci. USA* **93**, 7279–7282 (1996).
- Peretz, D. *et al.* A conformational transition at the N-terminus of the prion protein features in formation of the scrapie isoform. *J. Mol. Biol.* **273**, 614–622 (1997).
- Williamson, R. A. *et al.* Mapping the prion protein using recombinant antibodies. *J. Virol.* **72**, 9413–9418 (1998).
- Leclerc, E. *et al.* Immobilized prion protein undergoes spontaneous rearrangement to a conformation having features in common with the infectious form. *EMBO J.* **20**, 1574–1584 (2001).
- Chuong, C.-M., McClain, D. A., Streit, P. & Edelman, G. M. Neural cell adhesion molecules in rodent brains isolated by monoclonal antibodies with cross-species reactivity. *Proc. Natl Acad. Sci. USA* **79**, 4234–4238 (1982).
- Butler, D. A. *et al.* Scrapie-infected murine neuroblastoma cells produce protease-resistant prion proteins. *J. Virol.* **62**, 1558–1564 (1988).
- Caughey, B. & Raymond, G. J. Sulfated polyanion inhibition of scrapie-associated PrP accumulation in cultured cells. *J. Virol.* **67**, 643–650 (1993).
- Chabry, J. *et al.* Species-independent inhibition of abnormal prion protein (PrP) formation by a peptide containing a conserved PrP sequence. *J. Virol.* **73**, 6245–6250 (1999).
- Perrier, V. *et al.* Mimicking dominant negative inhibition of prion replication through structure-based drug design. *Proc. Natl Acad. Sci. USA* **97**, 6073–6078 (2000).
- Caughey, W. S., Raymond, L. D., Horiuchi, M. & Caughey, B. Inhibition of protease-resistant prion protein formation by porphyrins and phthalocyanines. *Proc. Natl Acad. Sci. USA* **95**, 12117–12122 (1998).
- Supattapone, S., Nguyen, H. O., Cohen, F. E., Prusiner, S. B. & Scott, M. R. Elimination of prions by branched polyamines and implications for therapeutics. *Proc. Natl Acad. Sci. USA* **96**, 14529–14534 (1999).
- Borchelt, D. R., Scott, M., Taraboulos, A., Stahl, N. & Prusiner, S. B. Scrapie and cellular prion proteins differ in their kinetics of synthesis and topology in cultured cells. *J. Cell Biol.* **110**, 743–752 (1990).
- Kaneko, K. *et al.* Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation. *Proc. Natl Acad. Sci. USA* **94**, 10069–10074 (1998).
- Zulianello, L. *et al.* Dominant-negative inhibition of prion formation diminished by deletion mutagenesis of the prion protein. *J. Virol.* **74**, 4351–4360 (2000).
- Scott, M. *et al.* Propagation of prions with artificial properties in transgenic mice expressing chimeric PrP genes. *Cell* **73**, 979–988 (1993).
- Priola, S. A. & Chesebro, B. A single hamster PrP amino acid blocks conversion to protease-resistant PrP in scrapie-infected mouse neuroblastoma cells. *J. Virol.* **69**, 7754–7758 (1995).
- Kocisko, D. A. *et al.* Species specificity in the cell-free conversion of prion protein to protease-resistant forms: A model for the scrapie species barrier. *Proc. Natl Acad. Sci. USA* **92**, 3923–3927 (1995).
- Priola, S. A., Chabry, J. & Chan, K. Efficient conversion of normal prion protein (PrP) by abnormal hamster PrP is determined by homology at amino acid residue 155. *J. Virol.* **75**, 4673–4680 (2001).
- Supattapone, S. *et al.* Prion protein of 106 residues creates an artificial transmission barrier for prion replication in transgenic mice. *Cell* **96**, 869–878 (1999).
- Xu, D. *et al.* In vitro characterization of five humanized OKT3 effector function variant antibodies. *Cell Immunol.* **200**, 16–26 (2000).
- Idusogie, E. E. *et al.* Mapping of the C1q binding site on rituxan, a chimeric antibody with a human IgG1 Fc. *J. Immunol.* **164**, 4178–4184 (2000).
- Schenk, D. *et al.* Immunization with amyloid-β attenuates Alzheimer-disease-like pathology in the PDAPP mouse. *Nature* **400**, 173–177 (1999).
- Bard, F. *et al.* Peripherally administered antibodies against amyloid β-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nature Med.* **6**, 916–919 (2000).
- Race, R. E., Fadness, L. H. & Chesebro, B. Characterization of scrapie infection in mouse neuroblastoma cells. *J. Gen. Virol.* **68**, 1391–1399 (1987).
- James, T. L. *et al.* Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl Acad. Sci. USA* **94**, 10086–10091 (1997).
- Rudd, P. M., Wormald, M. R., Wing, D. R., Prusiner, S. B. & Dwek, R. A. Prion glycoprotein: Structure, dynamics, and roles for the sugars. *Biochemistry* **40**, 3759–3766 (2001).

29. Petrescu, A. J., Petrescu, S. M., Dwek, R. A. & Wormald, M. R. A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology* 9, 343–352 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. Safar, S. DeArmond, F. Cohen and P. Dazin for discussions and K. Crossin for anti-N-CAM antibody. This work was supported by grants from the National Institutes of Health to R.A.W., D.R.B. and S.B.P.

Correspondence and requests for materials should be addressed to R.A.W. (e-mail: anthony@scripps.edu) or D.R.B. (e-mail: burton@scripps.edu).

Iron deficiency induces the formation of an antenna ring around trimeric photosystem I in cyanobacteria

Thomas S. Bibby, Jon Nield & James Barber

Wolfson Laboratories, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, London SW7 2AY, UK

Although iron is the fourth most abundant element in the Earth's crust, its concentration in the aquatic ecosystems—particularly the open oceans—is sufficiently low to limit photosynthetic activity and phytoplankton growth^{1,2}. Cyanobacteria, a major class of phytoplankton, respond to iron deficiency by expressing the 'iron-stress-induced' gene, *isiA* (ref. 3). The protein encoded by this gene has an amino-acid sequence that shows significant homology with one of the chlorophyll *a*-binding proteins (CP43) of photosystem II (PSII)^{4,5}. The precise function of the CP43-like protein, here called CP43', has not been elucidated, although there have been many suggestions^{3,6}. Here we show that CP43' associates with photosystem I (PSI) to form a complex that consists of a ring of 18 CP43' molecules around a PSI trimer. This significantly increases the size of the light-harvesting system of PSI. The utilization of a PSII-like protein as an extra antenna for PSI emphasises the flexibility of cyanobacterial light-harvesting systems, and seems to be a strategy which compensates for the lowering of phycobilisome and PSI levels in response to iron deficiency.

The cyanobacterium, *Synechocystis* PCC 6803, with a histidine (His)-tag attached to the carboxy terminus of the PSII chlorophyll *a*-binding protein, CP47 (ref. 7), was grown photoheterotrophically in the presence and absence of iron in the culture medium. PSI fractions free of PSII (see Methods) were resolved by sucrose density centrifugation. In the case of normal cells, the sucrose gradient separated two chlorophyll *a*-containing fractions, corresponding to PSI monomers and PSI trimers. In addition to these fractions, the iron-stressed cells produced two extra bands, one less dense than the PSI monomer and one more dense than the PSI trimer. SDS-polyacrylamide gel electrophoresis (SDS-PAGE; see Methods) showed (Fig. 1) the less dense of the iron-stress-induced fractions to be free CP43' (confirmed by immunoblotting). This technique also showed the more dense fraction to be a CP43'–PSI complex—consisting of the PsaA/PsaB reaction centre subunits of PSI (PSI RC, with a relative molecular mass of 75,000; M_r 75K), the CP43' protein and other lower- M_r PSI proteins. HPLC size exclusion chromatography gave an estimated M_r of approximately 1,900K for the CP43'–PSI complex, compared with 1,000K for the PSI trimer.

In Fig. 2 we show room-temperature optical absorption spectra, and 77K emission spectra, of PSI trimers, free CP43' and the

CP43'–PSI complexes. Although the absorption spectrum of the CP43'–PSI complex had a blue-shifted, long-wavelength absorption maximum compared with that of the PSI trimer due to the presence of CP43', the 77K fluorescence spectrum of the CP43'–PSI complex gave the same maximum as that of the PSI trimer alone. As isolated CP43' was highly fluorescent with a maximum at 685 nm,

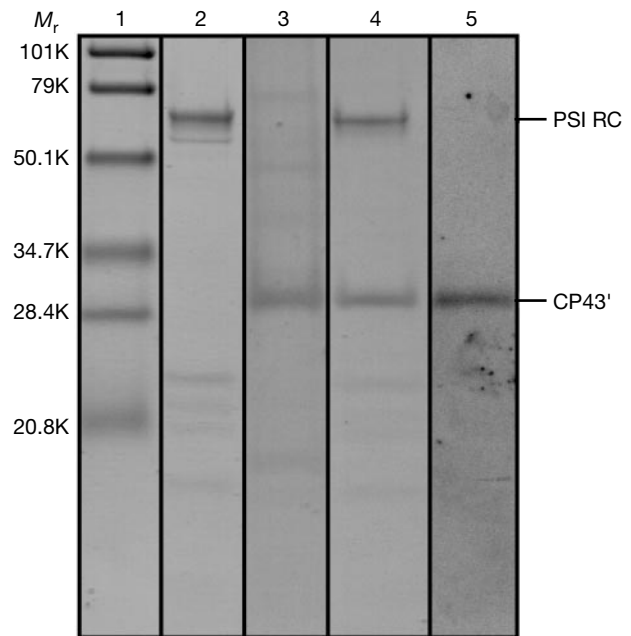


Figure 1 SDS-PAGE of fractions derived from sucrose density centrifugation. Lane 1, molecular markers; lane 2, PSI trimers; lane 3, CP43'; lane 4, CP43'–PSI complex; and lane 5, western blot using antibody to CP43'. PSI RC corresponds to the PsaA and PsaB reaction centre (RC) proteins of PSI. CP43' is the product of the *isiA* gene predicted to have M_r 37K but resolved at approximately 30K by SDS-PAGE.

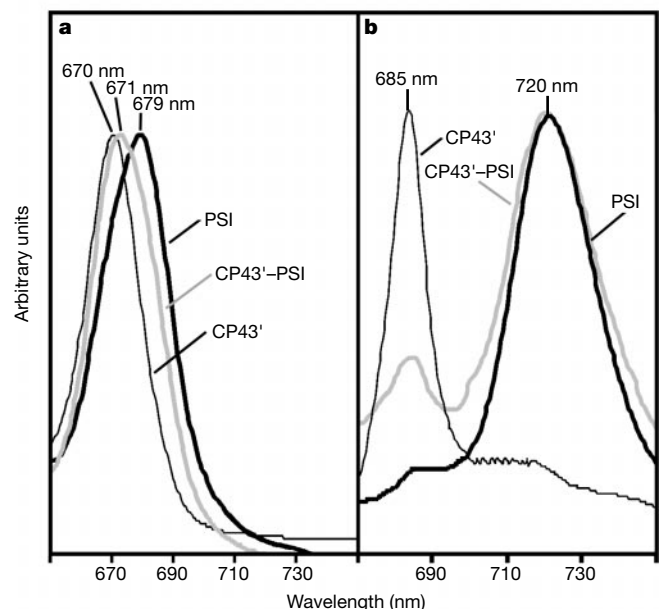


Figure 2 Optical absorption and fluorescence properties of fractions derived from sucrose density centrifugation. **a**, Room-temperature absorption spectra. **b**, Fluorescence emission spectra at 77K excited by 440-nm light which is absorbed by the Soret band of chlorophyll *a*. Spectra are normalized at their peaks to aid comparison.

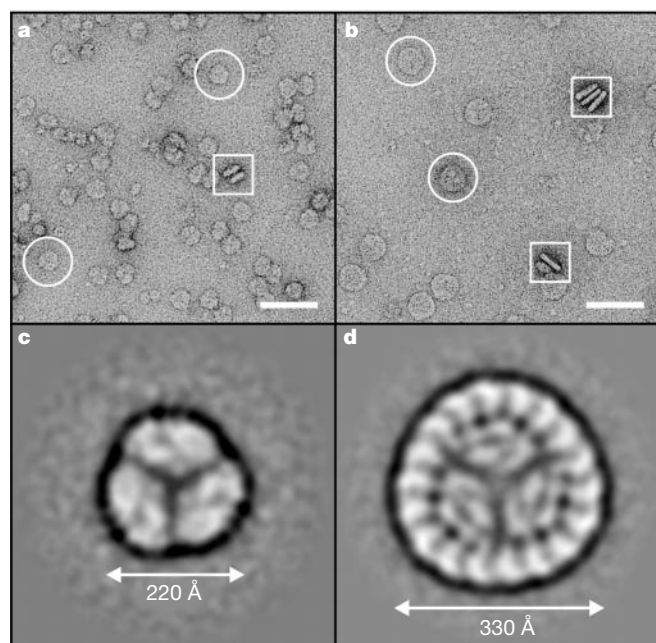


Figure 3 Electron micrographs and processed top views of the PSI trimer and CP43'-PSI supercomplex. Shown are typical electron micrographs of PSI trimers (**a**) and CP43'-PSI complex (**b**) in negative stain; scale bar, 50 nm. Most complexes are top views, probably of the stromal surface (ringed), but occasional side views were observed (boxed). Also shown are image-processed top views of negatively stained PSI trimers (**c**) and CP43'-PSI complexes (**d**). The projection maps were obtained by averaging 4,000 PSI trimer and 3,000 CP43'-PSI particles. Note that the iron-stress-induced complex contains a trimer of PSI surrounded by a 55 Å-thick ring of density corresponding to 18 copies of CP43'. The PSI trimer in (**c**) is surrounded by a detergent layer that accounts for its slightly larger size compared with the trimer within the CP43'-PSI complex. Similarly, the 55 Å ring of the latter complex also contains an outer detergent shell.

we conclude that in the CP43'-PSI complex, CP43' is associated with PSI in such a way as to allow efficient energy transfer. The small level of fluorescence at 685 nm from the CP43'-PSI complex seen in Fig. 2b can be attributed to some unattached CP43' in the preparation. Excitation spectra for the 720-nm emission from the CP43'-PSI complex confirmed that most of the CP43' present efficiently transferred energy to PSI.

To gain a better understanding of the association of CP43' with PSI, we conducted single-particle analysis of the CP43'-PSI complex on images obtained by electron microscopy. Figure 3a and b show typical electron micrographs of PSI trimers and the novel CP43'-PSI complex stained in uranyl acetate. In both cases, mainly top views were seen, although some stacked side views were present. Top-view averages of the PSI trimers and the CP43'-PSI complex are shown in Fig. 3c and d at a resolution of about 25 Å. The PSI trimer (Fig. 3c) corresponds in size and organization to that seen previously by electron microscopy^{8,9} and is consistent with X-ray diffraction studies on the PSI trimer of the thermophilic cyanobacterium *Synechococcus elongatus*¹⁰. The CP43'-PSI complex shown in Fig. 3d is more circular, with a diameter of about 330 Å corresponding to a 55 Å ring of CP43' subunits around the PSI trimer. The thicknesses of the PSI trimer and CP43'-PSI trimer supercomplex are approximately the same, as seen in Fig. 3a and b, and estimated to be about 90 Å. This indicates that the additional ring of CP43' in the CP43'-PSI complex is contained within the membrane. Image processing suggested that the ring was composed of 18 copies of CP43', and that the top views were probably of the stromal surface^{9,10}.

The 4 Å X-ray structure of the cyanobacterial PSI trimer has been determined¹⁰. Moreover, there is a 3.8 Å structural model of PSII

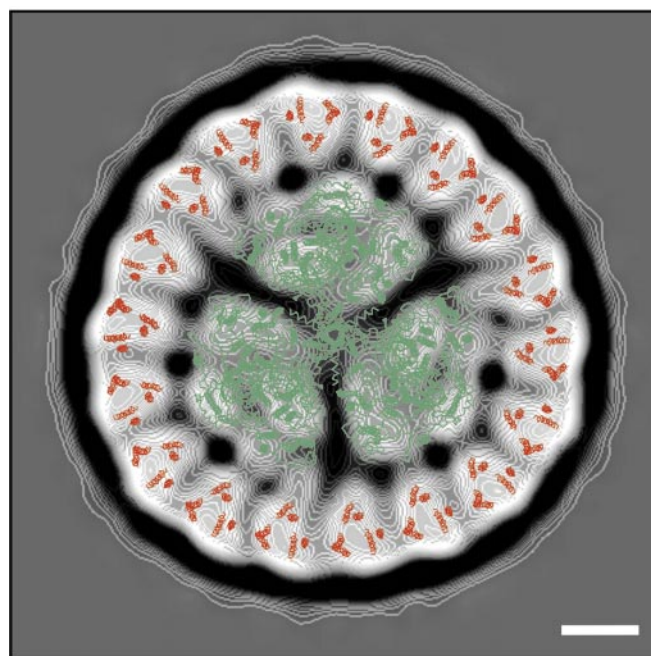


Figure 4 The PSI trimer structure and CP43 helix organization derived from X-ray crystallography^{10,11} overlaid onto the projection map of the CP43'-PSI complex. Scale bar, 5 nm. Good correlation between the X-ray and electron microscopy data confirms the presence of a PSI trimer within the centre of an 18-member ring of CP43'.

isolated from *S. elongatus*¹¹, which, as suggested from electron crystallography¹², shows CP43 to be a ring of three pairs of transmembrane helices. In Fig. 4 we have modelled the X-ray structures of the PSI trimer and CP43 into the projection map of the CP43'-PSI trimer supercomplex. This modelling suggests that there are indeed 18 copies of CP43' per supercomplex. Given that CP43 binds at least 12 chlorophyll molecules within its six-helix bundle¹¹, we conclude that the additional antenna size of the iron-stress-induced CP43'-PSI supercomplex is composed of approximately 216 chlorophyll *a* molecules. As the PSI trimer contains about 300 chlorophyll molecules¹³, then the light-harvesting ability of the supercomplex has increased by approximately 72% compared with that of the normal trimer. This increase in antenna size is almost certainly in response to the reduction in the level of the light-harvesting phycobilisomes and PSI complexes due to the heavy demand on iron for their synthesis and assembly³. Among the various hypotheses for the physiological role of CP43', it has been suggested that it may act as an additional light-harvesting system for PSII given that CP43' is a PSII-like protein³. Its functional association with PSII is therefore surprising. The formation of a ring of light-harvesting chlorophyll binding proteins around a reaction centre is reminiscent of the organization of the antenna systems of anoxygenic purple photosynthetic bacteria^{14,15}. We have shown that similar rings can occur in oxygenic photosynthetic organisms. □

Methods

Growth conditions

Synechocystis sp. PCC 6803 (ref. 7) was grown photoheterotrophically in mineral medium BG-11 (ref. 16), containing kanamycin and glucose at 30 °C, and with illumination of 70 μmol photons m⁻² s⁻¹. Iron-stressed cultures were obtained by inoculating the same BG-11 medium lacking iron-containing compounds. Cells were harvested after 3 d, and thylakoid membranes isolated according to a method similar to that described in ref. 17. The thylakoid membranes (1 mg chl ml⁻¹) were then solubilized with 1% β-D-dodecyl maltoside at 4 °C for 10 min. The solubilized membranes were centrifuged at 150,000g using a Beckman Ti70 rotor. The supernatant was then passed through a Ni²⁺-affinity column. PSII was selectively bound to the column via the His-tag while the non-bound fraction containing PSI was collected and subjected to sucrose density centrifugation.

Biochemical characterization

SDS-PAGE and western blotting were performed as described in ref. 18. Optical absorption spectra were measured at room temperature using a Shimadzu MPS 2000 spectrometer. Steady-state fluorescence spectra were obtained using a Perkin Elmer LS50 at 77K and measured with an excitation wavelength of 440 nm. To record excitation spectra, the sample was excited between 650 and 700 nm and the fluorescence detected at 720 nm. HPLC size exclusion analyses were performed using a Phenomenex BioSep SEC S3000 column.

Electron microscopy

Preparations were negatively stained with 2% uranyl acetate on glow discharged carbon evaporated grids, and imaged using a Philips CM 100 electron microscope at 80 kV. The magnification was calibrated to be $\times 51,500$. Twenty electron micrographs were taken for each preparation, and subsequently calculated to have the first minima of contrast transfer function in the range of 17–23 Å.

Image processing and modelling

Electron micrographs were digitized using a Leafscan 45 densitometer set at a step size of 10 μm . Single-particle data sets of approximately 3,000 (CP43'-PSI supercomplex) and 4,000 (PSI trimer) were obtained by interactively selecting all possible particles from the micrographs. All subsequent processing was performed within the IMAGIC-5 software environment^{19,20}. The single-particle images were coarsened by a factor of 2 resulting in 3.88 Å per pixel on the specimen scale. Reference free alignment coupled with multi-variate statistical classification²¹ was used to identify initial class averages, which were then used for iterative refinement until the data merged, resulting in the improved class averages shown.

Co-ordinate data sets were obtained from the RCSB data bank (<http://www.rcsb.org>) under the entry codes 1C51 (PSI 4 Å structure¹⁰) and 1FE1 (PSII 3.8 Å structure¹¹). These structural models were visualized using the program Swiss-PDB viewer²² (Glaxo-Wellcome Experimental Research) and overlaid at the same scale onto the calculated single-particle projection maps. The carbon- α backbone for the transmembrane helices of the CP43 subunit was extracted from the 1FE1 co-ordinates and modelled into each subunit of the ring surrounding the PSI trimer, according to the centre of mass observed for each of the 18 subunits within the ring.

Received 16 May; accepted 21 June 2001.

- Martin, J. H. *et al.* Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. *Nature* **371**, 123–129 (1994).
- Behrenfeld, M. J., Bale, A. J., Zbigniew, S. K., Aiken, J. & Falkowski, P. G. Confirmation of the iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean. *Nature* **383**, 508–511 (1996).
- Straus, N. A. in *Molecular Biology of Cyanobacteria* (ed. Bryant, D. A.) 731–750 (Kluwer Academic, Dordrecht, 1994).
- Burnap, R. L., Troyan, T. & Sherman, L. A. The highly abundant chlorophyll-protein complex of iron-deficient *Synechococcus* sp. PCC7942 (CP43') is encoded by the *isiA* gene. *Plant Physiol.* **103**, 893–902 (1993).
- Falk, S., Samson, G., Bruce, D., Hunter, N. P. A. & Laudenbach, D. A. Functional analysis of the iron-stress induced CP43' polypeptide of PSII in the cyanobacterium *Synechococcus* sp. PCC 7942. *Photosynth. Res.* **45**, 51–60 (1995).
- Park, Y. P., Sandström, S., Gustafsson, P. & Öquist, G. Expression of the *isiA* gene is essential for the survival of the cyanobacterium *Synechococcus* sp. PCC 7942 by protecting photosystem II from excess light under iron limitation. *Mol. Microbiol.* **32**, 123–129 (1999).
- Bricker, T. M., Morvant, J., Marsi, N., Sutton, H. M. & Frankel, L. K. Isolation of a highly active photosystem II preparation from *Synechocystis* 6803 using a histidine-tagged mutant of CP47. *Biochim. Biophys. Acta* **1409**, 50–57 (1998).
- Rögner, M., Mühlenhoff, U., Boekema, E. J. & Witt, H. T. Mono-, di- and trimeric PSI reaction center complexes isolated from the thermophilic cyanobacterium *Synechococcus* sp. Size, shape and activity. *Biochim. Biophys. Acta* **1015**, 415–424 (1989).
- Boekema, E. J., Boonstra, A. F., Dekker, J. P. & Rögner, M. Electron microscopic structural analysis of photosystem I, photosystem II, and the cytochrome *b₆f* complex from green plants and cyanobacteria. *J. Bioenerg. Biomembr.* **26**, 17–29 (1994).
- Krauß, N. *et al.* Photosystem I at 4 Å resolution represents the first structural model of a joint photosynthetic reaction centre and core antenna system. *Nature Struct. Biol.* **3**, 965–973 (1996).
- Zouni, A. *et al.* Crystal structure of photosystem II from *Synechococcus elongatus* at 3.8 Å resolution. *Nature* **409**, 739–743 (2001).
- Hankamer, B., Morris, E. P. & Barber, J. Revealing the structure of the oxygen evolving core dimer of photosystem II by cryoelectron crystallography. *Nature Struct. Biol.* **6**, 560–564 (1999).
- Schubert, W.-D. *et al.* Photosystem I of *Synechococcus elongatus* at 4 Å resolution: Comprehensive structural analysis. *J. Mol. Biol.* **272**, 741–769 (1997).
- McDermott, G. *et al.* Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **374**, 517–521 (1995).
- Cogdell, R. J. *et al.* How photosynthetic bacteria harvest solar energy. *J. Bacteriol.* **181**, 3869–3879 (1999).
- Williams, J. G. K. Construction of specific mutants in PSII photosynthetic reaction centres by genetic engineering methods in *Synechocystis* 6803. *Methods Enzymol.* **167**, 766–778 (1988).
- Tang, X.-S. & Diner, B. A. Biochemical and spectroscopic characterisation of a new oxygen-evolving photosystem II core complex from the cyanobacterium *Synechocystis* PCC 6803. *Biochemistry* **33**, 4594–4603 (1994).
- Hankamer, B. *et al.* Isolation and biochemical characterisation of monomeric and dimeric photosystem II complexes from spinach and their relevance to the organisation of photosystem II *in vivo*. *Eur. J. Biochem.* **243**, 422–429 (1997).

- van Heel, M., Harauz, G. & Orlova, E. V. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
- van Heel, M. *et al.* Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.* **33**, 307–369 (2000).
- Sherman, M., Soejima, T., Chui, W. & van Heel, M. Multi-variate analysis of single unit cells in electron crystallography. *Ultramicroscopy* **74**, 179–199 (1998).
- Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).

Acknowledgements

We thank A. Telfer, B. Hankamer, E. Morris, C. Büchel, J. Duncan and P. Da Fonseca for discussions; T. Bricker for supplying the His-tagged *Synechocystis* PCC 6803 mutant; and L. Sherman for supplying the CP43' antibody. This work was supported by the Biotechnology and Biological Research Council.

Correspondence and requests for materials should be addressed to J.B. (email: j.barber@ic.ac.uk).

A giant chlorophyll-protein complex induced by iron deficiency in cyanobacteria

E. J. Boekema*, A. Hifney†, A. E. Yakushevskaya*, M. Piotrowski‡, W. Keegstra*, S. Berry§, K.-P. Michel†, E. K. Pistorius† & J. Kruijs§

* Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands
† Biologie VIII: Zellphysiologie, Universität Bielefeld, D-33501, Bielefeld, Germany

‡ Plant Physiology, Ruhr-University Bochum, D-44780 Bochum, Germany
§ Plant Biochemistry, Ruhr-University Bochum, D-44780 Bochum, Germany

Cyanobacteria are abundant throughout most of the world's water bodies and contribute significantly to global primary productivity through oxygenic photosynthesis. This reaction is catalysed by two membrane-bound protein complexes, photosystem I (PSI) and photosystem II (PSII), which both contain chlorophyll-binding subunits functioning as an internal antenna¹. In addition, phycobilisomes act as peripheral antenna systems, but no additional light-harvesting systems have been found under normal growth conditions. Iron deficiency, which is often the limiting factor for cyanobacterial growth in aquatic ecosystems², leads to the induction of additional proteins such as *IsiA* (ref. 3). Although *IsiA* has been implicated in chlorophyll storage, energy absorption and protection against excessive light, its precise molecular function and association to other proteins is unknown. Here we report the purification of a specific PSI–*IsiA* supercomplex, which is abundant under conditions of iron limitation. Electron microscopy shows that this supercomplex consists of trimeric PSI surrounded by a closed ring of 18 *IsiA* proteins binding around 180 chlorophyll molecules. We provide a structural characterization of an additional chlorophyll-containing, membrane-integral antenna in a cyanobacterial photosystem.

We examined the protein composition of thylakoid membranes from the mesophilic cyanobacterium *Synechococcus* sp. PCC 7942 when cells were grown under conditions of iron limitation. Iron deficiency leads to fast degradation of phycobilisomes⁴ and to the expression of iron-regulated genes that enable cells to continue growth⁵. We found a distinct new protein complex in elution profiles of solubilized membrane protein complexes from iron-deficient cells (Fig. 1a). At the same time the amount of trimeric PSI, the most abundant protein complex under normal growth conditions, was reduced by 60–80%. The abundance of the new complex was inversely related to the disappearance of trimeric PSI

and represents an example of membrane-protein dynamics. We purified this complex to homogeneity by using an additional perfusion chromatography step (data not shown). The purified complex had an absorption maximum at 675 nm, resembling PSII (674 nm) rather than PSI (679 nm). Size-exclusion chromatography gave an apparent relative molecular mass of 1,700,000 (M_r 1,700K) for this complex (Fig. 1b). This is much larger than the molecular mass of 900K found for trimeric PSI. Analysis of subunit composition of the new complex by SDS–polyacrylamide gel electrophoresis (PAGE) showed that all major PSI subunits were present (Fig. 1c); however, the complex showed additional protein bands with apparent masses of 36, 48 and 53K (depending on the preparation). Obviously, the complex contains several copies of IsiA that are not always completely denatured and are separated by SDS–PAGE as several bands at higher molecular masses. *De novo* sequencing of tryptic peptides obtained from these bands resulted in perfect matches with the IsiA protein of *Synechococcus* sp. PCC 7942.

IsiA was first identified as a principal protein expressed under iron deficiency⁶ and is co-transcribed with flavodoxin, an alternative electron acceptor of PSI (ref. 7). The sequence of IsiA shows strong homologies to the CP43 subunit of PSII. The main difference is the lack of a large hydrophilic loop of about 100 amino acids. The structure of CP43 has been determined to 3.8 Å resolution as part of the dimeric cyanobacterial PSII core complex⁸. CP43 consists of six membrane-spanning α -helices to which at least 12 chlorophyll *a* molecules are bound⁸. As the chlorophyll-binding sites are conserved, a similar chlorophyll content of 12 can be estimated for IsiA. If one assumes an antenna with a chlorophyll content of 96 for PSI, a theoretical antenna size of 168 chlorophyll/P700 is reached for the PSI–IsiA complex (see below for the number of IsiA proteins per

PSI). This fits our chlorophyll/P700 measurements, which indicate a 60% increase in antenna size for the PSI–IsiA complex (161 ± 20 chlorophyll/P700) compared with trimeric PSI (103 ± 9 chlorophyll/P700). It was widely assumed that IsiA would interact with PSII (ref. 9) owing to its sequence homologies with CP43. IsiA (also called CP43') was proposed to protect PSII against excess light¹⁰, function as an alternative antenna for PSII (ref. 9), or act as a chlorophyll-storage protein⁵. In contrast, our data show that IsiA interacts with PSI. Qualitatively, the light-saturation curves of PSI and the PSI–IsiA complex look similar (Fig. 2), but a quantitative analysis using the equation:

$$\Delta A = \Delta A_{\max} \times [1 - \exp(-k \times \text{relative flash intensity})] \quad (1)$$

reveals that the absorption cross-section of the PSI–IsiA complex is about 44% higher than for PSI: *k*, which is a measure of the efficiency of photon capture, is 0.045 ± 0.003 for PSI and 0.065 ± 0.007 for the PSI–IsiA complex; *A* is the absorbance. This indicates, together with fluorescence spectra and PSI kinetics (data not shown), that IsiA is indeed a functional antenna for PSI. Furthermore, the determined structure of trimeric PSI at 2.5 Å resolution shows chlorophyll molecules that are bound to the outside of PSI (ref. 11). These pigments are ideally positioned for excitation energy transfer from the IsiA ring to the PSI reaction centre.

Imaging of the purified PSI–IsiA complex by electron microscopy revealed large numbers of a circular flat particle with a diameter of 345 Å (Fig. 3a). A data set of top-view projections was analysed by single-particle analysis, which revealed that the new complex is composed of a central trimeric PSI complex surrounded by a ring of 18 densities. Classification of projections indicated that the main variation in the data set was the amount of tilting. Well preserved, three-fold rotational symmetry, which is to be expected from zero-tilted particles, was only present in limited numbers of projections (Fig. 3b), indicating that about 95% of the particles are slightly tilted (Fig. 3c, d). The resolution in the images of Fig. 3b–d is about 22, 14 and 15 Å, respectively, allowing a close interpretation of the projected structure.

First, comparison with previously studied PSI particles indicated that the PSI–IsiA complexes^{12,13} are attached to the carbon support film by the bulky stromal surface. Apparently this 40-Å-high stromal ridge and the rather flat ring of peripheral IsiA subunits together cause the slight tilt of most of the particles.

Second, the peripheral IsiA ring can be further interpreted by

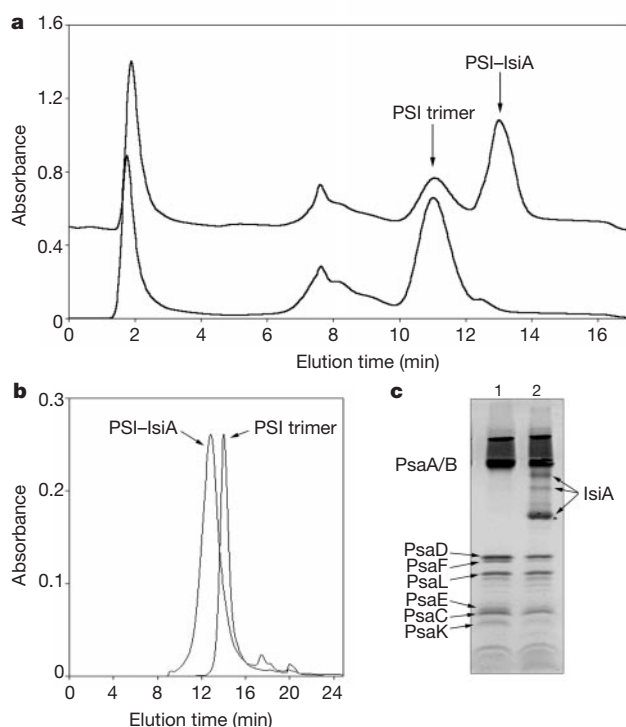


Figure 1 Purification and subunit analysis of the PSI–IsiA complex. **a**, Elution profiles for thylakoid membrane protein extracts of cells grown under either normal (lower trace) or iron-deficient conditions (upper trace). Purification was carried out on an anion-exchange column using a linear magnesium sulphate gradient for elution, and monitored at 435 nm. **b**, Size-exclusion chromatography of purified trimeric PSI and of PSI–IsiA complexes as indicated. **c**, SDS–PAGE and Coomassie staining of trimeric PSI (lane 1) and the PSI–IsiA complex (lane 2). Protein bands other than PSI subunits were assigned by *de novo* sequencing.

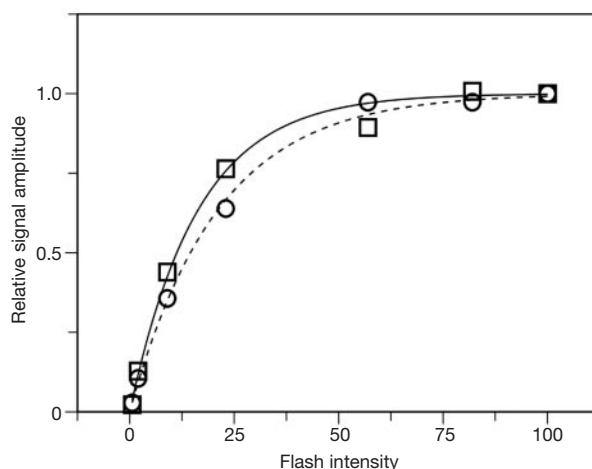


Figure 2 Light-saturation curves of the absorbance changes of PSI (circles and dashed line) and the PSI–IsiA complex (squares and solid line). Flash intensity was varied by a set of neutral-density filters. The absolute values of the signals at 100% flash intensity are $\Delta A = (2.5 \pm 0.2) \times 10^{-3}$ ($n = 7$) for PSI and $\Delta A = (1.6 \pm 0.2) \times 10^{-3}$ ($n = 6$) for PSI–IsiA. The curves are based on equation (1).

comparison to the homologous CP43 subunit of PSII (ref. 8). Its projected density has been modelled into the untilted projection (Fig. 3e) and fits into each of the 18 peripheral densities if the periphery is corrected for an attached detergent shell¹⁴. From this fit, it can be concluded that the new complex is composed of a ring of 18 IsiA proteins with a diameter of 310 Å. The PSI trimer is not a perfect circle and hence the 18 IsiA proteins do not form a perfect ring either. The two IsiA molecules that bind closest to the centre of each PSI monomer are shifted in an outward direction by about 7 Å and thus have a wider separation.

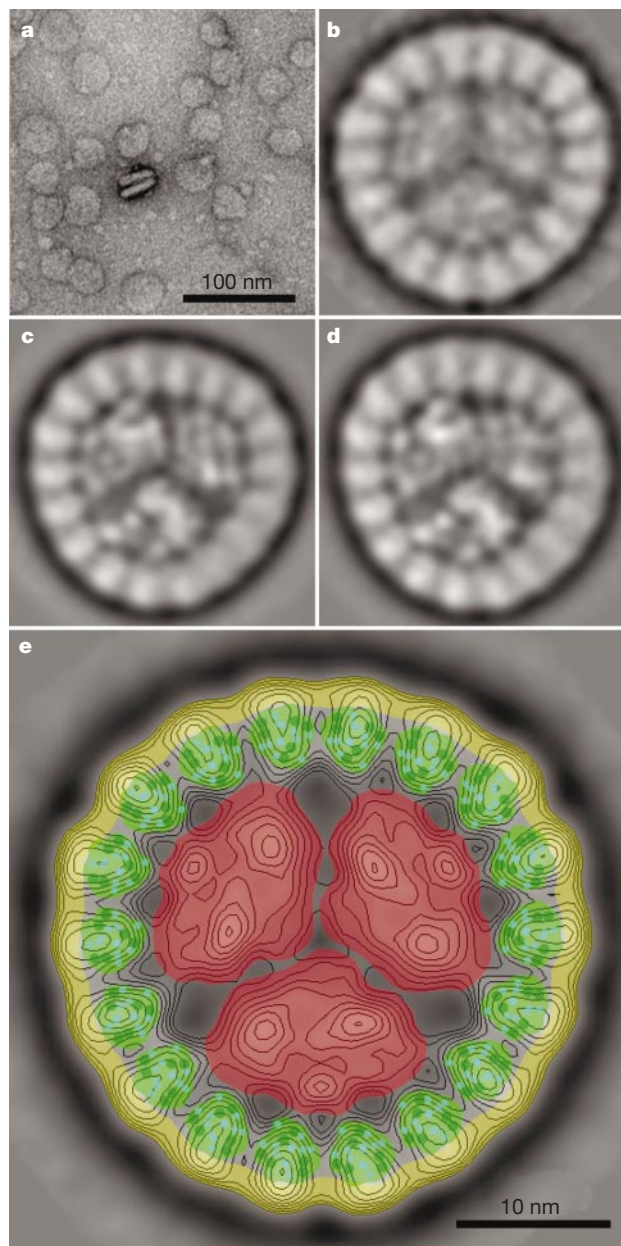


Figure 3 Electron microscopy of PSI–IsiA complexes and their interpretation. **a**, Electron micrograph of isolated complexes, negatively stained with uranyl acetate. **b–d**, Results of multivariate statistical analysis and classification of top-view projections; average images of 231 (**b**), 2,000 (**c**) and 2,000 (**d**) projections. **e**, Contoured version of the three-fold-symmetrical class sum from **b**. The centre of the complex is formed by a trimeric PSI complex (red) and is surrounded by a ring of 18 IsiA proteins (green), each consisting of six membrane-spanning α -helices (dark green dots) and 12 chlorophyll *a* molecules (blue dots). A detergent shell (yellow) surrounds the complex in projection. The images are presented as facing from the *p*-side of the thylakoid membrane, which is equivalent to the luminal side in green plants.

The gene coding for IsiA can be found in various cyanobacterial species where it is present in a single copy. Prochlorophytes, oxygenic prokaryotes possessing chlorophyll *b*, contain so-called *pcb* genes, which bind chlorophyll *b* and resemble IsiA (ref. 15). It will be interesting to discover how light-harvesting complexes are organized in this class of cyanobacteria, especially as it has been shown that prochlorophytes that are adapted to low light conditions contain several variations of *pcb* genes¹⁶. Our data give a clear structural characterization of a photosystem complex from a cyanobacterium with a peripheral membrane-bound antenna. This PSI–IsiA supercomplex is markedly different from all peripheral membrane-bound antenna complexes observed in green plants, including PSI and PSII (refs 12, 17). The IsiA ring resembles multimers of light-harvesting proteins (LH1 and LH2) found in anoxygenic purple bacteria^{18,19}. We now know of three peripheral chlorophyll-binding antenna systems: LH rings in purple bacteria, IsiA rings in cyanobacteria and LHC trimers/dimers in chloroplasts. Despite similarities in organization, all are structurally different from each other and are examples of convergent evolution. □

Methods

Cell culture

Synechococcus sp. PCC 7942 cells were grown in liquid BG-11 medium. Iron deficiency was achieved by omitting iron sources in the medium; cells were grown for 4 days (once diluted) in this medium. Growth conditions were as described²⁰ except that cultures were illuminated in constant light with fluorescent light tubes (Sylvania Luxline ES 18 W) at a light intensity of 120 $\mu\text{E m}^{-2} \text{s}^{-1}$.

PSI–IsiA purification

We isolated the PSI–IsiA complex from *Synechococcus* as described²¹ with some modifications. Solubilized membrane protein complexes were purified by anion-exchange chromatography (Poros 50 HQ, Applied Biosystems) using a magnesium sulphate gradient, followed by hydrophobic-interaction chromatography (Poros Butyl, Applied Biosystems) using an ammonium sulphate gradient. We used size-exclusion chromatography (BioSilect 400, Bio-Rad) as a final purification step before electron microscopy. Purified particles were concentrated by ultrafiltration (100K cut-off filter) and stored in buffer containing 20 mM MES, pH 6.5, 10 mM CaCl_2 , 10 mM MgCl_2 and 0.03% β -dodecyl maltoside at -70°C before use.

Mass spectrometry

Proteins were digested in gel using trypsin (Promega). Peptides were purified and concentrated using ZipTips C18 (Millipore). *De novo* sequencing was done on a quadrupole/time-of-flight hybrid mass spectrometer (Q-TOF2, Micromass) in positive-ion mode.

Electron microscopy

Transmission electron microscopy was performed with a Philips CM10 electron microscope at $\times 52,000$ magnification. We prepared negatively stained specimens on glow-discharged carbon-coated grids. From 119 digitized electron micrographs, 5,200 well preserved top-view projections were extracted for single-particle image analysis. Multi-reference alignment, multivariate statistical analysis and classification were performed with IMAGIC software^{22,23} and Grip software.

Difference spectroscopy

Flash-induced absorption changes at 703 nm were measured at a chlorophyll *a* concentration of 4 μM in a buffer containing 50 mM ACES, pH 6.5, 50 mM KCl, 30 μM phenazine methosulphate and 5 mM sodium ascorbate. Saturating blue light flashes were obtained by a xenon flash lamp equipped with blue glass filters. The antenna size was calculated from the signal amplitude using an extinction coefficient of 64 $\text{mM}^{-1} \text{cm}^{-1}$ (ref. 24).

Received 6 June; accepted 29 June 2001.

1. Nitschke, W. & Rutherford, A. Photosynthetic reaction centres: variations on a common structural theme? *Trends Biochem. Sci.* **16**, 241–245 (1991).
2. Behrenfeld, M., Bale, A., Kolber, Z., Aiken, J. & Falkowski, P. Confirmation of iron limitation of phytoplankton photosynthesis in the equatorial Pacific ocean. *Nature* **383**, 508–511 (1996).
3. Straus, N. A. (ed.) in *Advances in Photosynthesis* Vol. 1 *The Molecular Biology of Cyanobacteria* (ed. Bryant, D. A.) 731–750 (Kluwer, Dordrecht, 1994).
4. Guikema, J. & Sherman, L. A. Organization and function of chl in membranes of cyanobacteria during iron starvation. *Plant Physiol.* **73**, 250–256 (1983).
5. Burnap, R., Troyan, T. & Sherman, L. A. The highly abundant chlorophyll-protein complex of iron-deficient *Synechococcus* sp. PCC7942 (CP43') is encoded by the IsiA gene. *Plant Physiol.* **103**, 893–902 (1993).
6. Laudenbach, D. & Straus, N. A. Characterization of a cyanobacterial iron stress-induced gene similar to *psbC*. *J. Bacteriol.* **170**, 5018–5026 (1988).

7. Leonhardt, K. & Strauss, N. An iron stress operon involved in photosynthetic electron transport in the marine cyanobacterium *Synechococcus* sp. PCC 7002. *J. Gen. Microbiol.* **138**, 1613–1621 (1992).
8. Zouni, A. *et al.* Crystal structure of photosystem II from *Synechococcus elongatus* at 3.8 Å resolution. *Nature* **409**, 739–743 (2001).
9. Pakrasi, H. B., Riethman, H. C. & Sherman, L. A. Organization of pigment proteins in the photosystem II complex of the cyanobacterium *Anacystis nidulans* R2. *Proc. Natl Acad. Sci. USA* **82**, 6903–6907 (1985).
10. Park, Y.-I., Sandström, S., Gustafsson, P. & Öquist, G. Expression of the IsiA gene is essential for the survival of the cyanobacterium *Synechococcus* sp. PCC 7942 by protecting photosystem II from excess light under iron limitation. *Mol. Microbiol.* **32**, 123–129 (1999).
11. Jordan, P. *et al.* Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature* **411**, 909–917 (2001).
12. Boekema, E. J. *et al.* Green plant photosystem I binds light-harvesting complex I on one side of the complex. *Biochemistry* **40**, 1029–1036 (2001).
13. Lelong, C. *et al.* Characterization of a redox active cross-linked complex between cyanobacterial photosystem I and soluble ferredoxin. *EMBO J.* **15**, 2160–2168 (1996).
14. Dekker, J. P., Boekema, E. J., Witt, H. T. & Rögner, M. Refined purification and further characterization of oxygen-evolving and Tris-treated photosystem II particles from the thermophilic cyanobacterium *Synechococcus* sp. *Biochim. Biophys. Acta* **936**, 307–318 (1988).
15. La Roche, J. *et al.* Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. *Proc. Natl Acad. Sci. USA* **93**, 15244–15248 (1996).
16. Garczarek, L., Hess, W., Holtzendorff, J., van der Staay, G. & Partensky, F. Multiplication of antenna genes as a major adaptation to low light in a marine prokaryote. *Proc. Natl Acad. Sci. USA* **97**, 4098–4101 (2000).
17. Boekema, E. J. *et al.* Supramolecular organization of the photosystem II complex from green plants and cyanobacteria. *Proc. Natl Acad. Sci. USA* **92**, 175–179 (1995).
18. Karrasch, S., Bullough, P. & Ghosh, R. The 8.5 Å projection map of the light-harvesting complex I from *Rhodospirillum rubrum* reveals a ring composed of 16 subunits. *EMBO J.* **14**, 631–638 (1995).
19. McDermott, G. *et al.* Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **374**, 517–521 (1995).
20. Exss-Sonne, P., Toelle, J., Bader, K., Pistorius, E. & Michel, K.-P. The IdiA protein of *Synechococcus* sp. PCC 7942 functions in protecting the acceptor side of photosystem II under oxidative stress. *Photosynth. Res.* **63**, 145–157 (2000).
21. Wenk, S.-O. & Kruip, J. Novel, rapid purification of the membrane protein photosystem I by high-performance liquid chromatography on porous materials. *J. Chromatogr. B* **737**, 131–142 (2000).
22. Van Heel, M. & Frank, J. Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**, 187–194 (1981).
23. Haraux, G., Boekema, E. & van Heel, M. Statistical image analysis of electron micrographs of ribosomal subunits. *Methods Enzymol.* **164**, 35–49 (1988).

24. Hiyama, T. & Ke, B. Difference spectra and excitation coefficients of P700. *Biochim. Biophys. Acta* **267**, 160–171 (1972).

Acknowledgements

We thank A. Köhl and F. Pfeil for technical assistance, and our colleagues M. Rögner and A. Trebst for discussions. Financial support of the Deutsche Forschungsgemeinschaft (J.K. and E.K.P.), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (E.J.B.) and the Fonds der chemischen Industrie (J.K.) is gratefully acknowledged.

Correspondence and requests for materials should be addressed to E.K.P. (e-mail: e.pistorius@biologie.uni-bielefeld.de) or J.K. (e-mail: jochen.kruip@ruhr-uni-bochum.de).

erratum

Elevated c-myc expression facilitates the replication of SV40 DNA in human lymphoma cells

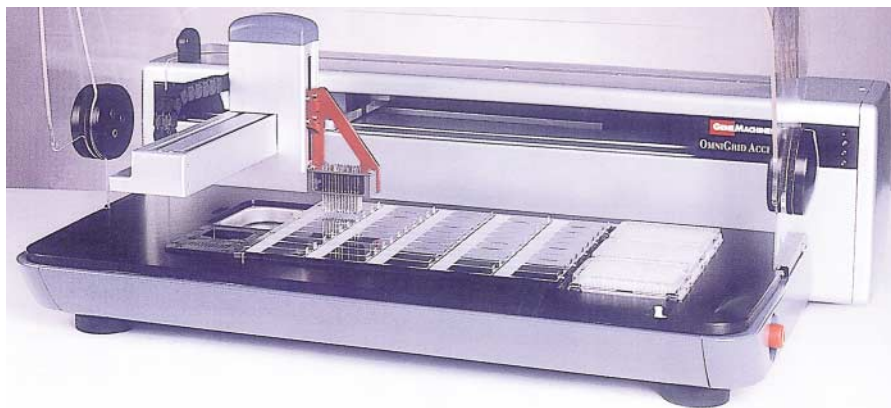
Marie Classon, Marie Henriksson, Janos Sümegi, George Klein & Marie-Louise Hammar skjöld

Nature **330**, 272–274 (1987).

In this Letter, the last name of Marie-Louise Hammar skjöld was misspelled as 'Hammaskjöld'. □

Peptides and proteins

Assay kits, centrifuges and other protein-friendly paraphernalia.



Bench-top microassays: the latest compact from OmniGrid.

OmniGrid Accent

GeneMachines www.genemachines.com
Mind your Accent

The compact OmniGrid Accent microarrayer prints from three-microtitre plates onto a 50-slide capacity platform and supports a print head holding up to 48 pins. Centre-to-centre spot distances can be adjusted in 2.5-micron increments. The arrayer prints over 4,000 features per slide per hour at full pin and slide capacities. Using the same software interface as the OmniGrid, the unit allows users to control and customize sample tracking, array formats, pin configurations and wash protocols that include any user-defined combination of sonication, water/fluid wash with peristaltic pump and vacuum dry. It is supplied with a clear-panelled enclosure and a positive-pressure HEPA filtration system for dust-free printing.

Tp-Blue/ Tp-Red

Biocolor www.biocolor.co.uk
Total protein assay kits

Advertised as 'next generation' assay kits, Tp-Blue Coomassie Brilliant Blue G has a linear range of 0.5–10 µg protein per 10 µl sample, and Tp-Red Pyrogallol Red-Molybdate produces a linear range of 5–50 µg protein per 10 µl sample. The dye-protein complex remains in solution for approximately 10 hours, and reagents are methanol-free. They are supplied in kits sufficient for 1,000 96-well assays or 4,000 384-well assays.

LightShift

Pierce www.piercenet.com
Chemiluminescent EMSA kit

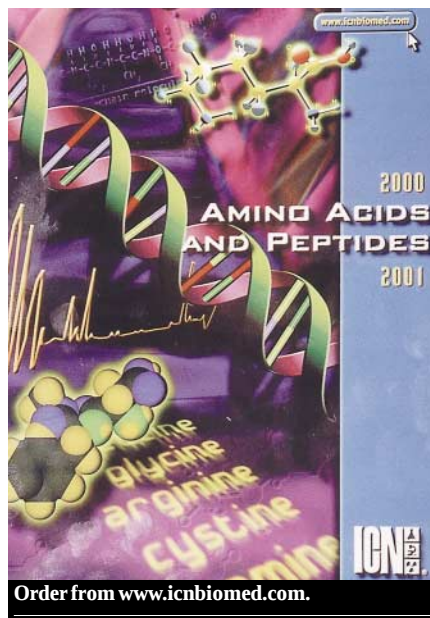
Often referred to as gel shift or gel retardation, the electrophoretic mobility shift assay (EMSA) is used to evaluate DNA-protein interactions. It is based on the principle that,

when subjected to electrophoresis, free DNA will migrate differently from a DNA-protein complex. LightShift is a non-radioactive system capable of sub-femtomole detection levels, offering the sensitivity and speed of radioactive gel shift assays without the hazards, waste and probe instability problems associated with radioactive systems. It is supplied with all necessary components for performing 100 DNA-protein binding reactions, with the exception of positively charged nylon membrane, and sufficient chemiluminescent detection reagents for 800 cm² of membrane.

Amino acids and peptides catalogue

ICN Biomedicals www.icnbiomed.com
A weighty tome

This free catalogue contains a comprehensive listing of amino acids, amino acid derivatives and biologically active peptides. Individual sec-



Order from www.icnbiomed.com.

tions include amino acids, *t*-BOC, CBZ and Fmoc-protected amino acids, biologically active peptides, substrates, inhibitors, synthesis reagents, and books. A technical section covers protecting group strategy in peptide synthesis and includes reference tables of molecular weights.

Proteasome reagents

Affinity BioReagents www.bioreagents.com
New range of polyclonal antibodies

Proteolytic degradation is crucial in maintaining appropriate levels of short-lived regulatory proteins. Thirteen new polyclonal antibodies to the different subunits of the proteasome complex have been introduced by Affinity BioReagents. Antibodies to both the 20S and 26S proteasomes and their subunits as well as proteasome activators, inhibitors I, II, III and IV and a large selection of ubiquitin products are available. All antibodies have been optimized for western blot applications in a wide range of species.

GeNet 2.3

Silicon Genetics www.sigenetics.com
New software release for gene expression data

GeNet 2.3 is the latest version of this company's web-enabled database. Fully compatible with GeneSpring 4.0, it distributes and visualizes gene expression data from any organism. Researchers can use it to publish and store experiments, gene lists, clusters, pathways and any other type of analysis performed from within GeneSpring. GeNet 2.3 features enhanced browser-based mining and administrative capabilities and a revamped GenePage that allows the user to view raw, normalized and control values of a chosen gene along with the gene annotation. GeNet can run off a flat-file back end (standard edition) or using Oracle database technology (enterprise edition).

Rapid-Screen and Rapid-Scan

Cambridge Bioscience www.bioscience.co.uk
Panels for cloning and expression

Rapid-Screen arrayed cDNA library panels are designed to rapidly and easily clone full-length cDNAs. To overcome the drawback of the relatively low average insert size inherent in most libraries, the Rapid-Screen library first isolates the DS cDNAs of different size fractions and then ligates them separately into the vector. The use of 96-well plate arrays and a PCR-based screening procedure allows clones to be identified in three rounds of PCR. Vector and gene-specific primers can be used in concert to determine the size of cDNA



Centurion: the Bio-Plex protein array analyses up to 100 proteins and peptides from one sample.

Bio-Plex protein array

Bio-Rad

www.bio-rad.com

A protein array system to conserve samples

The new protein array system allows the simultaneous multiplexed analysis of up to 100 different proteins and peptides from as little as 12 μ l of sample. The integrated system includes immunoassay panels in 96-well format based around the current version of the Luminex 100 fluorescent analyser. The Bio-Plex assay panels focus on cell signalling pathways, providing information of up to 100 proteins. Assays are available for cytokines and phosphoproteins from both mouse and human serum, and a custom assay service is also provided.

ZipTipHPL

Millipore

www.millipore.com

Wash away those blues

These pipette tips contain hydrophilic interaction chromatography resin to remove detergents and Coomassie Blue stain from peptide mixtures. When used prior to mass spectrometry they can improve data quality. The tip is placed on a single- or multichannel pipettor or compatible automated liquid handling station. Sample is bound by aspirating and dispensing through the micro-volume resin bed. Detergents or stain are washed away and the concentrated, purified sample is eluted.

Proteome analysis

BioCat GmbH www.biocat.de/ www.igb.fhg.de

Do it your way

BioCat and Fraunhofer IGB are cooperating to provide customized transcriptome and proteome analysis. For each custom proteomics project performed by Fraunhofer IGB, the customer receives a detailed design and analysis report. The modules on offer are sample preparation, comparative 1D or 2D protein gel electrophoresis, comparative quantitative analysis of electrophoresis results, and cutting out of bands or spots of interest. Contact BioCat for a quote on a specific problem.

These notes are compiled in the Nature office from information provided by the manufacturers.

inserts after the initial PCR screen, eliminating the need to isolate many clones in order to determine which are full length. Rapid-Scan gene expression panels are used to determine in which tissues a gene is expressed and to survey the level of expression. Up to 24 tissues can be examined simultaneously, with quantitative expression data obtained in around three hours. This compares favourably with northern blot techniques in which only 6–12 tissues can be examined at any one time.

Microfuge 22R

Beckman Coulter

www.beckmancoulter.com

Five go spinning

The benchtop Microfuge 22R is designed for rapid pelleting of DNA and proteins. It spins up to thirty 1.5-ml tubes at 21,290g while maintaining a temperature of 4°C at maximum speed, for sample protection. The unit accommodates five different rotors, all of which feature aerosol-tight lids to minimize contamination from broken or leaking tubes. The rotors are autoclavable at 121°C.

QIAexpress

Qiagen

www.qiagen.com

Vectors for His tags

QIAexpress is a new trio of vectors for the expression of His-tagged proteins. Vector pQE-30 UA, contained in the QIAexpress UA cloning kit, is a prelinearized pQE-30 UA expression vector into which PCR products generated using non-proofreading DNA polymerases such as *Taq* can be inserted. Stringent and efficient ligation is achieved by the use of 3'-end U overhangs on the vector and 3'-end A overhangs on the PCR product, together with an optimized ligation master mix. There is no need for restriction digestion of the vector or insert. Vector pQE-TriSystem allows His-tagged protein expression in the three most commonly used expression systems without the need for subcloning procedures. Parallel expression of a His-tagged protein in *E. coli*, baculovirus-infected insect cells and mammalian cells can be achieved using a single construct. Vector pQE-30Xa is supplied with Fac-

tor Xa protease and Xa removal resin. Together they can be used to remove His tags from recombinant proteins to deliver the target protein free of vector-encoded amino acids. The insertion of a factor Xa protease recognition site between the 6 × His tag and the first amino acid of a target protein allows efficient cleavage of the His tag by treatment with factor Xa protease, which is subsequently removed using Xa removal resin.



It's cold in here: Microfuge 22R.

ADVERTISEMENTS

Polyclonal Antibodies & Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

Proteasome inhibition?

~ epoxomicin ~
~ PR39 ~ lactacystin ~
~ clasto-lactacystin ~ β ~
~ lactone ~ P131 ~
~ MG-132 ~ MG-262 ~

www.proteasome.com

AFINITI Research Products Ltd.
Mamhead Castle, Exeter, EX6 8HD, U.K.
T.+44(0) 1626 891010 – F+44(0) 1626 891090
– Email info@proteasome.com –

Contacts

Publisher: Fabien Savenay

Editor: Paul Smaglik

Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:
Leonie Welss (4954)

European Manager:
Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayouni (4978)

Scandinavia: Silke Opstrup (4994)

Production Manager: Billie Franklin

To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Aminda Wren
Tel + 34 91 391 3835
Fax + 34 91 319 0091
e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: classified@natureny.com

US Sales Director: Ben Crowe

US Sales Manager: Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Shinjuku-Ku,
Tokyo 162
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejp.com
Japan Manager: Kate Cowan

Mixed fortunes in France

Employment trends, no matter how tenuously discerned or documented, are inevitably good news for some groups and bad for others. This point hit home during a recent visit to Paris. In conversations with officials from both the European Synchrotron Radiation Facility and the Pasteur Institute, spokespeople from each organization noted that they were having trouble attracting postdocs. After further probing, both officials backtracked a little. They were getting plenty of applicants, they said. The problem was in attracting quality candidates.

Both provided plenty of informed speculation as to why this is so — at least, as informed as my attempts to play amateur sociologist and derive some sort of meaning from this small sample size. The good economy meant more permanent positions — in academia and industry — for newly minted PhDs, they said. And they noted that permanent positions are desirable, not just for job security, but also for the ability to carry out research unencumbered by somewhat arbitrary deadlines.

Of course, I can sympathize with organizations wanting to attract the 'quality' candidates. It is only natural that they should seek the best possible scientists. But their present discouragement may have more to do with high expectations of what, exactly, a PhD should bring to a postdoc position.

Should they be highly trained scientists who can carry out complex research tasks for a principal investigator? Or should they be promising, but somewhat inexperienced researchers who, with training, can thrive after a few years? The French postdoc shortage is bad for the relatively few institutions seeking the former, but good for the relatively many young scientists aspiring to the latter.

Paul Smaglik
Naturejobs editor



Contents

MOVERS

Physicist downsizes; US institute gets new director; synchrotron expert takes on Diamond; and more [Back page](#)

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

FINANCE

Ann Simon, a physicist who has made a career of monitoring biotechnology companies, will later this year move from a mid-size biotech firm to a smaller one. Simon, formerly group finance director of Bioglan Pharma, has been appointed finance director of Adprotech. At Bioglan, Simon enjoyed taking the company public and swelling its coffers through a secondary offering. She looks forward to growing the smaller company in a similar manner. "My personal preference is for smaller companies where there is a much wider range of skills involved in the finance-director role," she says. Before her three years at Bioglan, Simon spent 13 years in corporate finance at Cazenove, latterly specializing in biotechnology and science-based companies. Simon, who joined 'the city' "because it looked like fun", notes that finance is a good match for trained scientists. "Many of the skills you have in academia, particularly project management, ability to multitask and ability to clarify and solve problems, are directly transferable into either the corporate world or the corporate advice world," she says.

US NATIONAL INSTITUTES OF HEALTH

Biochemist **Lutz Birnbaumer**, a cell biologist specializing in how cellular signalling regulates organ and muscle function, was named scientific director of the US National Institute of Environmental Health Sciences (NIEHS) last month. Birnbaumer is professor and chair of the department of molecular, cell and developmental biology at the University of California, Los Angeles, professor of anaesthesiology and biological chemistry, and a full member of the university's Institute of Molecular Biology. In October, Birnbaumer will replace **J. Carl Barrett**, who held the NIEHS position for 23 years. **Paul Nettiessheim**, who is set to retire as director of the NIEHS laboratory of pulmonary pathobiology, has served as acting scientific director in the interim.

ECOLOGY AND CONSERVATION

Richard Brusca, the interim director of education at Columbia University's Biosphere 2 Center, was last month appointed director of conservation and science at the Arizona-Sonora Desert Museum. He will lead the museum's research and conservation programmes as well as oversee its

long-term conservation policies. Brusca has led field research expeditions ranging from projects in Antarctica to the Sonora Desert. He succeeds **Gary Nabhan**, who accepted the director's chair for the newly established Center for Sustainable Environments at Northern Arizona University. Nabhan will continue his research affiliation with the museum.

BIOLOGY

The Stowers Institute for Medical Research in Kansas City, Missouri, continues its recruitment drive, last month adding three new researchers. **Brian Sauer**, who developed the Cre-Lox recombinase technology, is joining as the institute's new director of transgenic technology in September. He comes from the Oklahoma Medical Research Foundation, although he developed the Cre-Lox system, a widely used gene-knockout tool, while at E. I. du Pont de Nemours and DuPont Merck Pharmaceutical from 1984 to 1993. **Scott Hawley**, a new Stowers senior scientist, from the University of California, Davis, and **Ranjan Perera**, the institute's associate director of genomics, from Akkadix in La Jolla, California, both joined the institute last month.

PHYSICS

Gerhard Materlik, associate director of the HASYLAB synchrotron radiation facility in Hamburg, will run Diamond, a third-generation synchrotron facility being built at the Rutherford Appleton Laboratory near Oxford, Britain. Diamond is financed by the British and French governments along with the Wellcome Trust. Materlik helped to build HASYLAB and served as its director from 1986 to 1993. He has also been involved in the European Synchrotron Radiation Facility from its inception in 1978 and has served on the advisory boards for synchrotrons around the world, including Spring-8 in Japan and the Advanced Photon Source in the United States. Diamond is scheduled to open in September 2006 and will include an initial seven beamlines with space for a further 37.

Erratum

On page 11 of the 21 June issue of *Naturejobs*, a caption identifying Kirsten Drejer, president of Symphogen is incorrect. The photo is of Kirstin Kriz, of the European Institute of Science.



Ann Simon



Richard Brusca



Brian Sauer

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.